2010

Efficient processing of system scenarios in statistical and machine learning studies for power system operational and investment planning

Venkat Kumar Krishnan
Iowa State University

Efficient processing of system scenarios in statistical and machine learning studies for power system operational and investment planning

by

Venkat Kumar Krishnan

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Electrical Engineering

Program of Study Committee:
James D. McCalley, Major Professor
Venkataramana Ajjarapu
Manimaran Govindarasu
Mervyn Marasinghe
Sigurdur Olafsson

Iowa State University

Ames, Iowa

2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and gratitude to my advisor Prof. James D. McCalley for his valuable guidance and timely encouragement throughout the course of my doctoral studies. My deepest gratitude is due to my POS committee members for their valuable time and comments.

I would like to gratefully acknowledge the support of French Transmission operating company RTE-France, especially Henry Sebastian and Samir Issad for their active participation as industry advisors through the course of this research work.

I wish to express my special thanks to my wife Trishna Das and mother Mrs. Usha Krishnan for their amazing support and sacrifices. I am deeply grateful to my Teacher to have selflessly extended The Supreme Lord’s causeless mercy in my life and bestowing all brightness in my otherwise darkened heart. My sincere prayer is to use all the gifts of Supreme Lord in His service.

I would also like to express my thanks to my friends Dr. Siddhartha Khaitan, Dr. Kasthurirangan Gopalakrisahnan, Dr. Amit Pande, Abhisek Mudgal, Sparsh Mittal, Sidharath Jain, Sandeep Krishnan, Dr. Ankit Agrawal, Dr. Sivakumar Swaminathan, Ganesh Ram Santhanam and Vikram Koundinya for truly making this whole graduate school experience a life transforming one.

ABSTRACT

Power System security assessment and the associated planning studies are becoming more and more complex with ever increasing uncertainties in all time horizons. An effective means of performing operational and investment planning studies of network limitations associated with static or dynamic post-disturbance performance problems has been to take a Monte Carlo simulation based approach. The approach harnesses computing power to develop a database of post-contingency response over a wide range of different operating conditions, and then apply statistical or machine learning methods to extract useful planning and operational information from the database.

Key to the machine learning based planning approach is the manner in which the different operating conditions are sampled to generate a training database. This work develops an efficient sampling procedure that maximizes information content in the training database while minimizing computing requirements to generate it, by finding the most influential region in the sampling state space and sampling operating conditions from it according to their relative likelihood. The Monte-Carlo variance-reduction methods are used to construct the proposed sampling approach, which is envisioned to allow market-oriented industries to operate the system according to economic rule.

The dissertation also develops a comprehensive methodology to perform decision tree based security assessment for multiple contingencies. The system security limits and associated operating rules depend on the set of contingencies considered for planning. Considering the probabilistic nature of the power system, this work develops a risk based contingency ranking method that helps in screening the most critical contingencies from a contingency list. The developed contingency risk estimation method gives realistic risk

indices since it takes into account the non-parametric nature of operating condition distribution, and it also saves tremendous computational cost since it uses linear sensitivities to estimate the risk. Finally, a contingency grouping method is proposed that guides in generating common operating rules for every group that performs well for all the contingencies in that respective group, thereby providing system operators the benefit of dealing with lesser number of rules. The contingency grouping is based on newly devised metric called *progressive entropy* that helps in finding similarities among contingencies based on their consequences on the operating conditions along all the load ranges, and not just their proximity in the grid.

The proposed methods are implemented in the west France, Brittany region of RTE-France's test system to derive decision rules for multiple contingencies against voltage stability problems.

CHAPTER 1 INTRODUCTION

1.1 INTRODUCTION

In the modern society, electric power is considered as one of the very vital commodities. With the growing dependence on industries in the current highly competitive economy and people's fast-paced life style, there is a great importance given to power system reliability assessments and planning. Traditionally such studies in power system involve deterministic assessment techniques and criteria, that are being used in practical applications even now, such as WECC/NERC disturbance-performance table for transmission planning [1, 2]. But the drawback with deterministic criteria is that they do not reflect the stochastic or probabilistic nature of the system in terms of load profiles, component availability, failures etc [3]. Furthermore, in the current market oriented power structure where heavy transactions are happening over long transmission lines in an interconnected environment, the system is constantly pushed to its stability limits, and the number of uncertainties has increased tremendously with respect to generation dispatch, reactive resource availability etc. Therefore the need to incorporate probabilistic or stochastic techniques to assess power system reliability and obtain suitable indices or guidelines for planning has been recognized by the power system managers, planners and operators; and several such techniques have been developed [4, 5, 6, 7, 8].

1.2 PROBABILISTIC RELIABILITY EVALUATION METHODS IN POWER SYSTEM

Power system reliability assessment can be divided into system adequacy (long term planning) and system security (operational) studies [9]. The term adequacy refers to the existence of sufficient resources to satisfy load entities or operational constraints, which

include facilities necessary to generate sufficient energy, reliably transport the energy produced to the load entities. The term security refers to the ability of the system to respond to dynamic or transient disturbances, which includes events such as contingencies that could lead to system instabilities etc.

Typically reliability evaluation techniques can be divided into two categories [9]:

- *Analytical:* Represent the power system using analytical models and evaluate the indices using mathematical solutions.
- *Simulation:* Monte Carlo simulation (MCS) methods used to estimate the indices or generate post-contingency data by simulating the actual process with randomness of system states.

MCS methods have several advantages such as [9]:

- Several system effects or process including nonelectrical factors such as weather effects etc. can be included in the study which may have to be approximated in analytical methods.
- They can simulate from the probability distributions of the parameters to be sampled such as component failure or system operating conditions etc.
- They can also provide probability distribution of performance measure random variables which have great practical significance.

An overview of simulation methodology is shown in Fig. 1.1. It involves two major tasks: *database generation approach* and *statistical or machine learning analysis* as illustrated by left-hand-side and right hand side of the figure respectively.

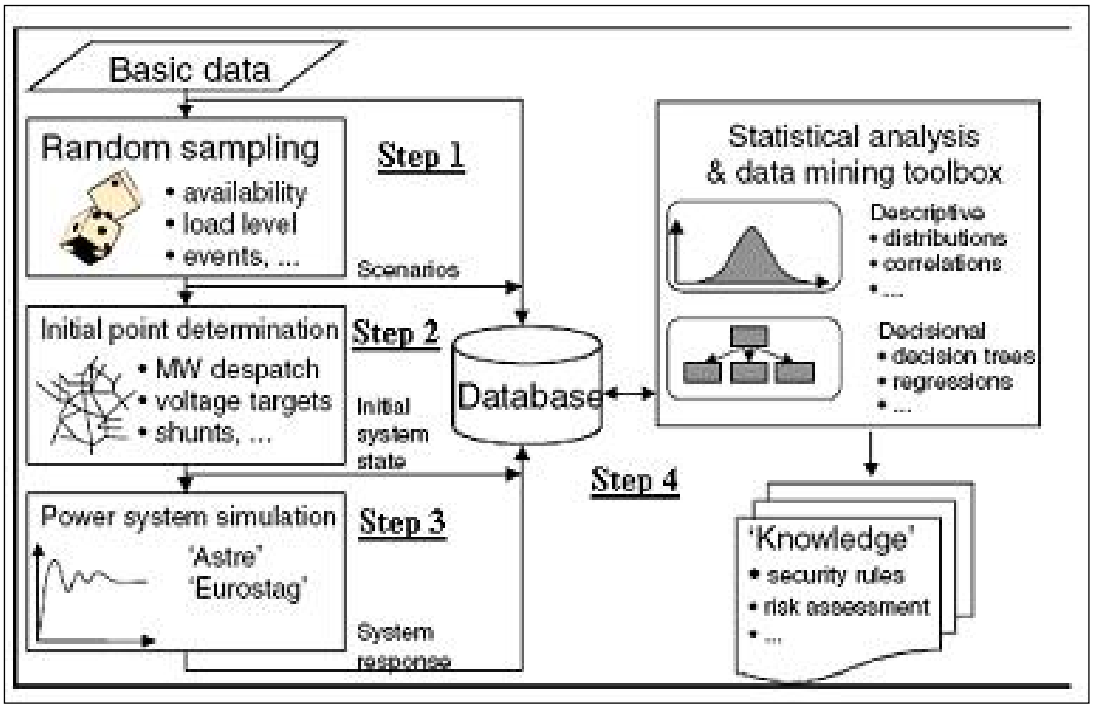


Fig. 1.1 Power system probabilistic reliability analysis overview [10]

1.2.1 Database Generation Approach

Database generation approach involves the following steps:

(1) *Random Sampling:* Operating parameters (load, unit commitment, circuit outages, power transfers etc.) are selected, assigned a distribution (e.g., uniform, Gaussian, exponential, empirical (historical records) etc.) and randomly sampled. This process is generally known as Monte Carlo sampling.

(2) O*ptimal power flow* run to obtain the initial state, and

(3) *Contingency events* are simulated using steady-state or time-domain (dynamic) simulation, and post-contingency performance measures are obtained.

1.2.2 Statistical Analysis

The object of many simulation experiments in power system is the estimation of an expectation $E[g(X)]$, where X is a random vector, typically the system performance measure

obtained from contingency analysis output of data generation step. The expectation functions estimated based on performance measure typically provides system reliability indices.

Such system reliability evaluation using MCS methods has been extensively developed in the domain of adequacy assessment [9, 11, 12] to evaluate:

(i) *Generating capacity reliability* with indices such as loss of load expectation (LOLE), Loss of energy expectation (LOLE) etc.

(ii) *Composite system reliability* with indices such as Expected load curtailments (ELC), Expected demand not served (EDNS), Expected energy not served (EENS) etc.,

(iii) *Distribution system reliability* with indices such as System average interruption frequency index (SAIFI), System average interruption duration index (SAIDI) etc.,

(iv) *Reliability worth/cost* with indices such as Expected interruption cost (EIC) etc.

For system security assessment studies, MCS is typically done to estimate risk-based system security limits with respect to transient stability, thermal overload, voltage stability etc., such as maximum allowable system loadability, expected ATC, expected voltage stability margin etc [13, 14, 15, 16].

1.2.3 Automatic Machine Learning Techniques

Automatic machine learning methods [17, 18], also known as knowledge discovery from databases, are used to extract a high level information, or knowledge from a huge database containing post-contingency responses obtained from database generation step. The machine learning or data mining techniques are broadly classified as:

- *Unsupervised learning:* Those methods which do not have a class or target attribute. For example, association rule mining can be used to find the correlation between various attributes. Clustering methods such as *k*-means, EM etc. are generally used to discover classes.
- *Supervised learning:* Those methods that have a class or target attribute, such as classification, numerical prediction etc., and use the other attributes (other observable variables) to classify or predict class values of scenarios. For example, naïve bayes, decision trees, instance based learning, neural network, support vector machine, regression etc.

With the increase in computing power, this tool has been widely used in many disciplines ranging from psychology, medical diagnosis, image-processing, and so on. In the field of power system, it has found a very great application in security assessment [19, 20, 21, 22, 23]. Other avenues of power system where they find application are design of protection systems, load forecasting [24], load modeling, state estimation, equipment monitoring etc.

1.2.3.1 Decision Tree Based Inductive Learning

There is particularly a great interest in using decision trees in power system security assessment for their ability to give explicit rules to system operators in terms of critical pre-contingency system attributes. These operating rules help in guiding operators in energy control centers as shown in Fig. 1.2, during conditions for which contingencies may result in violation of reliability criteria. So effectively these operating rules help operators map the pre-contingency scenarios to post-contingency consequences, thereby in a predictive fashion delineating secure operating regions from insecure operating regions in the space of pre-contingency parameters accessible in control centers such as flows, generation levels, load

levels etc. Therefore the proximity to a security boundary can be easily monitored, and when an alarm is encountered the operator must take appropriate control action to move into a more secure operating condition. This gives the operators a very simple and easy way to monitor and handle the power system operation, which otherwise is tedious for such a huge non-linear dynamic system.

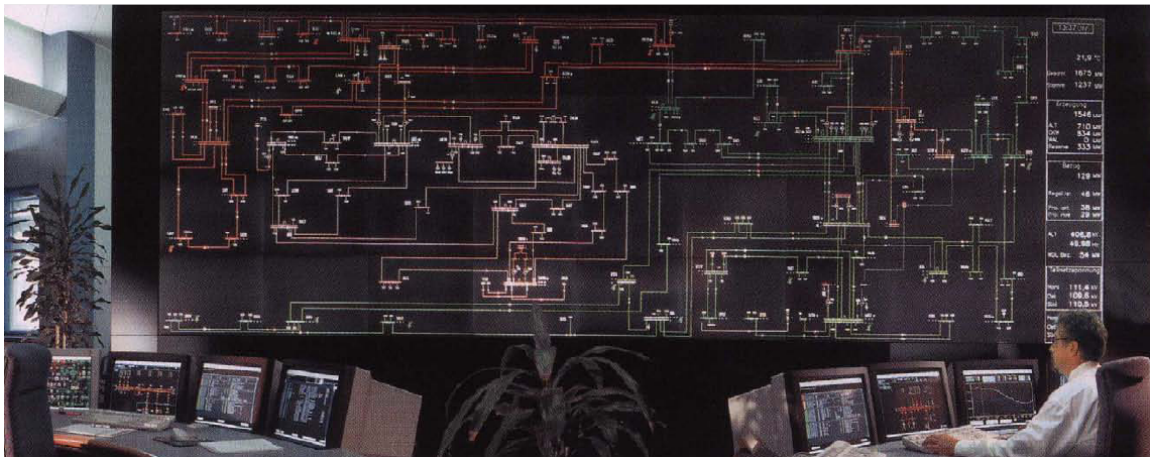


Fig. 1.2 Typical control center environment – Operational rules application

So the decision tree based inductive learning method enables decision-makers in an operational planning environment to establish operating guidelines or rules in terms of threshold values of various critical pre-contingency system attributes, in order to figure out the conditions of power system during which it is secure/stable from post-contingency performance point of view [25, 26, 27, 28, 29].

The inductive learning is performed on the database obtained from database generation step and operational rules are derived, which is deductively applied to learn unknown scenarios. Information required for building decision tree:

- A training set, containing several pre-contingency attributes with known class values

- The classification variable (i.e., class attribute with typical class values such as “secure” or “insecure”) could be based on post-contingency performance indices like voltage stability margin, etc.
- An optimal splitting rule, i.e., rule to find critical attribute
- A stopping rule, such as maximum tree length, depth, or minimum instances etc.

Basic Algorithm:

- **INPUT** the training/learning data into the topmost node
- **IF** stopping rule applies for the given input dataset, **THEN** stop, **ELSE** Apply the optimal splitting rule to select the best attribute for splitting the top node
- Using the splitting rule, decompose the learning set into *‘p’* mutually exclusive subsets. Usually p = 2, binary tree with two outcomes such as “secure” and “insecure”
- **IF** classification achieved (use stopping rule), **THEN** return classification, **ELSE** branch by setting ‘splitting’ attribute to each of the possible threshold values (can be interpreted as rules), and repeat with branch as your new tree, and the subset of data as the learning set

The aim is to obtain a model that classifies new instances well and produces simple to interpret rules. Ideally we would like to get the best model that has no diversity (impurity), i.e., all instances belong to same class. But due to many other uncertainties or interactions that have not been accounted for in the model, there would be some impurity (i.e., non-homogeneous branch) at most of the levels. So the goal is to select attributes at every level of branching such that impurity or diversity is reduced. There are many measures of impurity, which are generally used as optimal splitting criteria to select the best attribute for splitting. Some of those are Entropy, Information, Gini Index, Gain Ratio etc.

Classification accuracy and error rates can be used as the performance measures of a decision tree. There are two kinds of errors: *False Alarms* - Acceptable cases classified as Unacceptable; and *Risks* - Unacceptable cases as Acceptable. Errors can be calculated by testing the obtained decision model on the training set, which is usually an over-estimate. There are some training set sampling methods such as *holdout procedures, cross-validation, bootstrap etc* [18] to make the error estimation unbiased. It is even better if the testing is performed using an independent test dataset. Typically some portion of the original data is reserved for training and the remaining data used for testing. A rule of thumb is $1/3^{rd}$ for testing and $2/3^{rd}$ for training. There are numerous references [18] that explain the process of building a decision tree from a database with algorithms such as ID3, J48 etc. CART [30], Answer Tree [31], Orange [32], WEKA [33] etc. are some software available for building decision trees.

Many utilities have taken and are continuing to take a serious interest in implementing learning algorithm such as decision tree in their decision making environment. French transmission operator RTE has been using decision tree based security assessment methods to define operational security rules, especially regarding voltage collapse prevention [34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. They provide operators a better knowledge of the distance from instability for a post-contingency scenario in terms of pre-contingency system conditions, and they save a great amount of money while preserving the reliability of the system by enabling more informed control of the operation nearer to the stability limits.

1.2.4 Summary

Monte Carlo simulation based approach has been an effective means of performing operational and investment planning studies of network limitations associated with static or

dynamic post-disturbance performance problems. The approach harnesses computing power to develop a database of post-contingency response over a wide range of different operating conditions. Then statistical or machine learning methods such as decision tree are applied to extract useful knowledge from the database for decision making.

1.3 OBJECTIVES

This section presents the two major objectives of this dissertation work, along with the motivations behind the work and the significance of accomplishing the objectives.

1.3.1 Efficient Processing of System Scenarios

The most vital and sensitive part of Monte Carlo simulation based reliability studies is the stage of database generation. The confidence we will have in the results generally reflects the confidence we have in the set of system states generated. While in the case of statistical studies, the generated database does influence the quality of the estimate, in the case of machine learning studies it does influence the classification performance of the derived operating guidelines against realistic scenarios, selection of critical rule attributes and their threshold values (which will have bearing on economic operation of power system), and size of the operating rules. Furthermore, it typically incurs a high computational cost to improve the quality of estimates in statistical studies and rule's performance in machine learning studies. So there is this contradictory objective of increasing the information content or intelligence in the database generation step at the expense of minimal computational cost.

As mentioned, in the case of statistical studies the database generation stage using MCS methods typically become very time consuming as it needs very large sample size for estimating reliability indices with good accuracy (low variance). This is especially true for cases estimating reliability indices related to rare events. But this issue has been addressed

using several Monte Carlo variance reduction techniques, which have been applied in practice [45, 46, 47, 48, 49, 50] to improve the accuracy of estimation and also reduce computational cost. But in the field of decision tree based reliability assessment studies, the challenge of producing high information content training database at lower computational cost has not been addressed adequately [51, 52, 53, 54]. In the open literature, there are re-sampling methods to retain only the most important instances from an already generated training database [55, 56] for classification purposes. But such methods involve huge computational cost in first generating a training database, then identifying the most influential instances, and if need be, generate more of such instances. Recently, Genc et. al. [57] proposed an iterative method to a-priori identify the most influential region in the operating parameter state space, and then enrich the training database with more instances from the identified high information content region for enhancing classification performance. In this case, the method proposed to identify the high information content region involves heavy computational cost when the dimension of the operating parameter space increases, even beyond 10 parameters. Furthermore, the work doesn't delineate the tremendous advantages of training a decision tree using high information contained database, but rather waters down its significance by including training instances also from other regions, that may not be so influential to the decision making process.

So the primary objective of this dissertation is to develop an efficient database generation method that creates a satisfactory training database with low computational cost by sampling most influential operating conditions from the input operating parameter state space prior to the stage of power system contingency simulation. In short, the objective is to ***maximize information content in the training database, while minimizing computing requirements*** to

generate it. This work develops a linear sensitivity based method to very quickly identify the high information content region in a multidimensional operating parameter state space with non-parametric probability distribution. The work clearly explains and demonstrates the advantage of exclusively generating a training database from the identified high information content region of the operating parameter state space.

1.3.2 Decision Tree based Operational Planning for Multiple Contingencies

The reliability assessment, and consequently the short term operational and long term investment planning solution strategies depend on the set of contingencies considered in the planning study. Typically, a thorough contingency analysis of many contingencies is performed, and the most important ones based on system reliability limits are screened. Then appropriate solution strategies are devised, i.e., in our case relevant decision trees are developed to address every critical contingency screened.

In order to reduce the computational burden of contingency analysis, contingency ranking methods are typically used in power system reliability assessment studies. They help in fast screening of the most critical set of contingencies for thorough analysis and planning. While there are many deterministic ranking methods that considers the impact or severity of contingencies [58, 59]; under the current highly probabilistic nature of power system, a contingency ranking method which does not consider the probability of contingencies would lead to misleading operational solutions strategies against real time conditions. This is because the risk posed by a contingency under a wide variety of operating conditions not merely depends on its severity, but also on its probability of occurrence. So we propose to develop a ***risk based contingency ranking process*** that would eventually help in screening top contingencies

leading to voltage collapse, where the risk index of a contingency is estimated as the product its severity over various operating conditions and its probability.

The decision tree based operational planning for multiple contingencies is further advanced by the proposed concept of ***contingency grouping***. The proposed contingency grouping method will strike a balance between producing simple and accurate trees, as well as reducing the number of trees for multiple contingencies. The grouping of contingencies is based on a novel criterion, known as *progressive entropy curves*, that reflects the overlap among various contingency's effect on operating conditions, which is unlike traditional methods based on geographical proximity.

1.4 Dissertation Organization

The rest of this dissertation is organized as follows:

Chapter 2 presents the proposed efficient sampling strategy to generate database with high information content for training decision trees. The chapter gives a detailed description of the "information content" concept, and systematically presents the two-stage efficient training database generation method constructed using Monte Carlo variance reduction techniques. The efficient sampling approach developed is demonstrated on French EHV network to derive operating rules against voltage stability problems. The chapter also gives detailed account of extracting relevant historical data for our study from French SCADA system.

Chapter 3 addresses the very important issue of capturing finer details of multivariate load distribution such as its non-parametric nature and the mutual correlation in order to generate realistic operating conditions using the Monte Carlo sampling process. The chapter also focuses on the development of fast state space characterization method based on LHS of

stress directions and linear sensitivity measures. The developed efficient processing method is applied on French EHV network for security assessment against voltage stability problems, and the results are analyzed in great detail. The chapter also sheds some light on the simulation methodologies used to realize the fast characterization of parameter state space.

Chapter 4 presents a comprehensive security assessment method based on decision trees for multiple contingencies. With earlier chapters as the backbone to perform the security assessment, the crux of the chapter deals with two concepts to build the comprehensive multiple contingency security assessment process: risk based contingency ranking and contingency grouping. The chapter presents a detailed technical description of both the concepts, and presents the application results for seven contingencies considered in the west region of French network.

Chapter 5 presents conclusions and significant contributions of this work, and discusses possible future works.

CHAPTER 2 HIGH INFORMATION CONTENT DATABASE GENERATION FOR DECISION TREE BASED OPERATIONAL PLANNING

2.1 INTRODUCTION

Decision tree based inductive learning method serves as an attractive option for preventive-control approach in power system security assessment. It identifies key pre-contingency attributes that influence the post-contingency stability phenomena and provides the corresponding acceptable scenario thresholds. These guidelines are deductively applied to classify any new pre-contingency scenario with respect to its post-contingency performance, thereby enabling maximum utilization of available resources without compromising the reliability of power system in real time.

2.2 MOTIVATION AND PROPOSAL

Database generation for training is the critical aspect of performance of any data mining based power system reliability studies. Generally a uniform or random sampling of system states is carried out by varying parameters such as load level, unit commitment, system topology, exchanges at the boarders, component availability etc. according to their independent probability distributions obtained from projected historical data [10, 25, 38, 42, 43, 60] or forecasted 24-hour data [26, 27, 28, 29]. Then, various scenarios are simulated for a pre-specified set of contingencies or faults. This stage is generally very tedious and time consuming, as there could be a tremendously large number of combinations of variables and topologies, even within a ‘study region’ (about 5000-15000 samples for a statistically valid study [10]). Some studies [25, 26, 28] expend extra computation after validating the

operational rules to increase the unstable (rare) situations in database to improve the accuracy. While this would reduce one type of error, namely ‘risk’ of misclassifying unacceptable scenario as acceptable, it does not address the other error, namely ‘false alarm’ due to misclassifying acceptable scenario as unacceptable. Moreover if the sampled unstable situations are unrealistic or unlikely, then it could make the rules very conservative, i.e either costly to respect or sending irrelevant warning regarding the true limit of the system (more false alarms) by misclassifying acceptable scenarios as unacceptable.

In this chapter, we propose to develop an efficient sampling method to generate influential operating conditions that captures high information content for better classification and also reduces computing requirements. This efficient sampling is constructed using the Monte Carlo Variance Reduction (MCVR) techniques. Among the mostly used MCVR methods, control variate and antithetic variate take advantage of the correlation between certain random variables to obtain variance reduction in statistical estimation studies. Stratification method and importance sampling method re-orient the way the random numbers are generated, i.e., alters the sampling distribution [61, 62]. The proposed efficient sampling method is constructed using the importance sampling method for its ability to bias the Monte Carlo sampling towards the influential region identified a-priori; and generate samples within the influential region preserving the original relative likelihood of the operating conditions.

In order to sample the influential operating conditions, the operating parameter state space must be characterized with respect to post-contingency performance first. C. Singh et. al. [47] proposed a state space pruning method to identify the important region in a discrete parameter space composed of generation levels and transmission line capacities under a

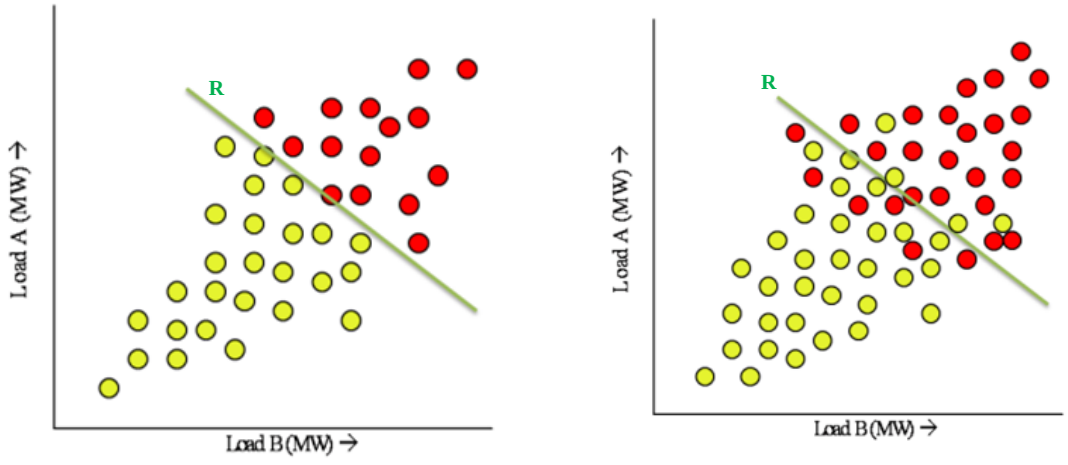
single load level for system adequacy assessment. X. Yu et. al. [63] proposed self-organized mapping, a unsupervised neural network, together with Monte Carlo simulation (MCS) to characterize the transmission line state space. The method that we have developed uses stratified sampling to characterize operational parameter state space.

The remaining parts of this chapter are organized as follows. Section 2.3 describes the concept “information content” in the context of this work. Section 2.4 presents the technical approach of the proposed high information contained training database generation. Section 2.5 demonstrates the application in deriving operational rules for voltage stability problem in Brittany region of RTE’s system, and presents results. Section 2.6 concludes.

2.3 High Information Content

The decision tree learning algorithm requires a database that has good representation of all the class values, so that it can effectively classify new instances and not overlook the less representative classes. So, for a two-class problem, a good representation of operating conditions on both sides of the class boundary is required. Also, not every operating condition on both sides of the class boundary contributes equally to the operating rule derivation process. For instance, consider sampling some operating conditions defined in terms of variations in Loads A and B as shown in Fig. 2.1a. Perform contingency analysis to find the post-contingency voltage stability performance. A suitable rule can be defined by line R that effectively partitions the operating region with acceptable post contingency performance from unacceptable performance. We refer to this line as the security boundary. Now, if more operating conditions are sampled as shown in Fig. 2.1b, the samples drawn near to the security boundary influences the rule making process more than the samples away from the boundary. This is evident from the consequent rule change (shifting line R) that is

necessary as shown in Fig. 2.1c. So it is very essential that the database contains operating conditions nearer to the security boundary with finer granularity, since they convey more information on the variability of the performance measure, which thereby enables a clear cut decision making on the acceptability of any operating condition. Furthermore, if the some of the operating conditions with unacceptable performance near the rule line R in Fig. 2.1c are less likely to occur in reality, then the rule line R may be shifted slightly upwards to exploit more operating conditions for economic reasons, as shown in Fig. 2.1d. Hence the desired influential operating conditions are obtained by sampling according to the probability distribution of the boundary region, which is the shaded region in Fig. 2.1d where there is a high uncertainty in the acceptability of any operating condition. This will also ensure a very good representation of both the classes in the database at a reduced computational cost compared to sampling from the entire operational parameter state space probability distribution.



(a) (b)

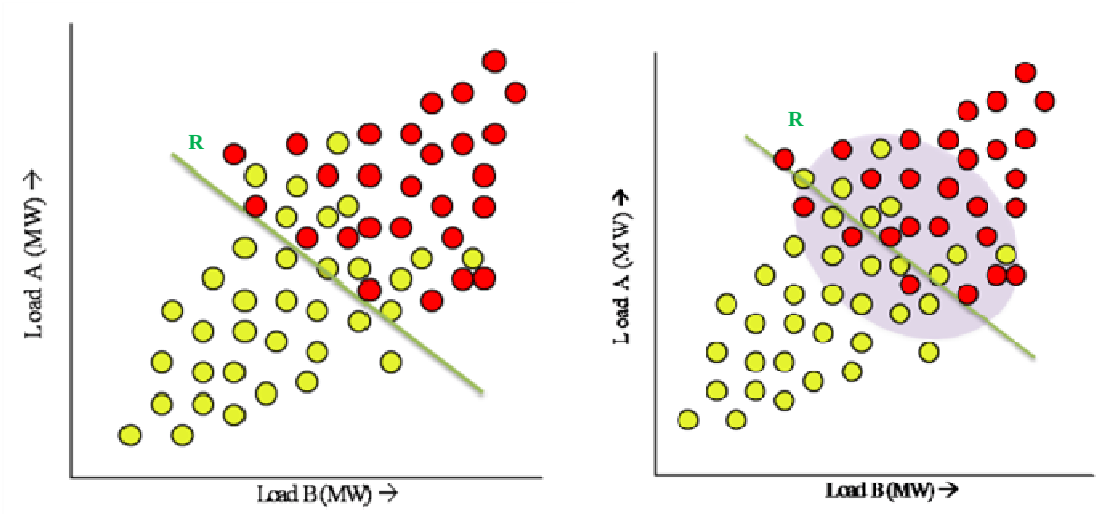


Fig. 2.1 High information content region

In this work *Entropy*, the most commonly used information theoretic measure for the information contained in a distribution, is used to quantify information content in a database [64]. It is a function of class proportions, when operating conditions are sampled according to their probability distribution. *Entropy* is given by equation (2.1)

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \qquad (2.1)$$

where, S is training data, c is the number of classes, and p_i is the proportion of S classified as class i. Given that the security boundary generally falls in the lower probability region of the operating parameter state space, a database containing samples within the boundary region has the maximum entropy, produced at reduced computational cost.

2.4 TECHNICAL APPROACH

The overall flowchart of risk-based planning approach is shown by Fig. 2.2, along with the proposed efficient sampling approach. The proposed algorithm consists of two stages, where stage I utilizes a form of stratified sampling to approximately identify the boundary region and stage II utilizes importance sampling to bias the sampling towards the boundary region.

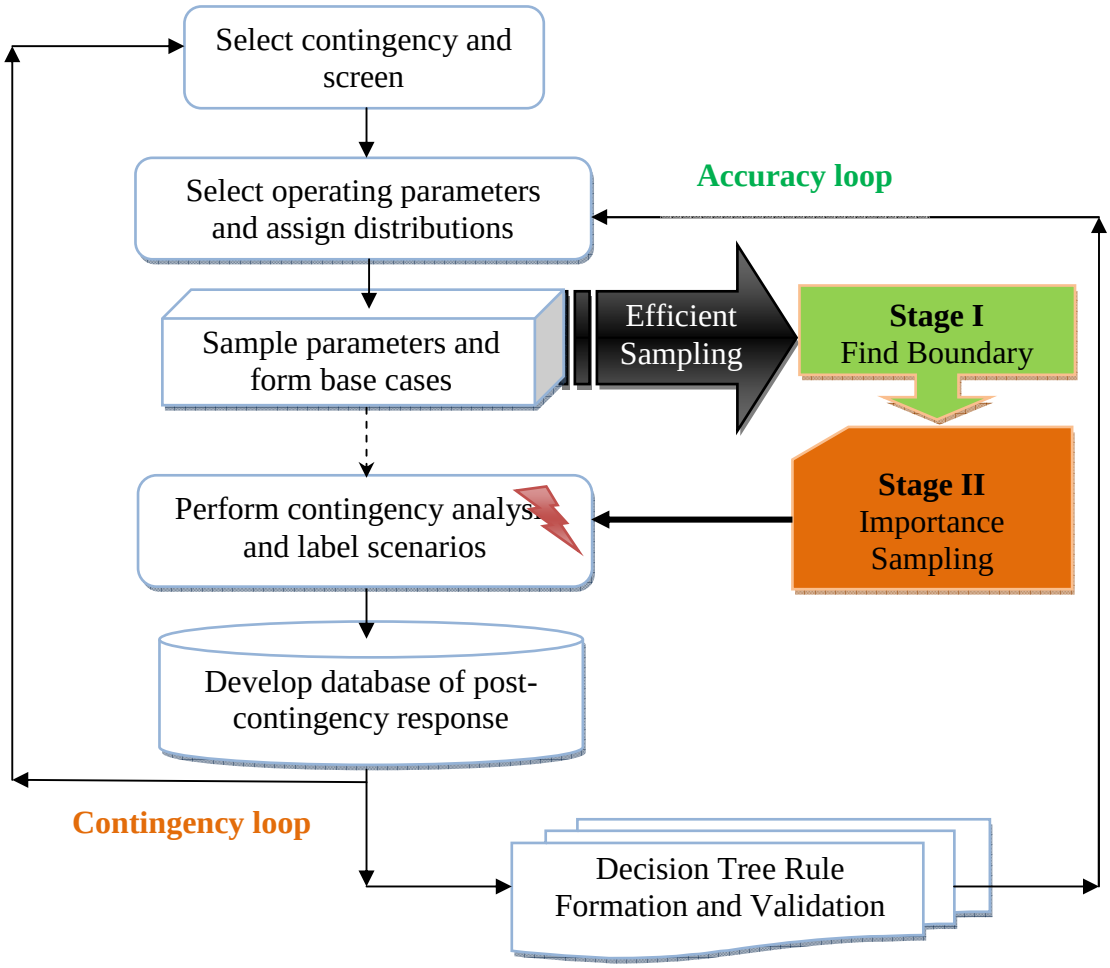


Fig. 2.2 Proposed approach

The database generation is performed for every critical contingency or a group of critical contingencies screened, as depicted by the contingency loop. The accuracy loop feeds back information about the region of sampling state space requiring more emphasis in training database, in order to reduce decision tree misclassifications and improve the accuracy. This chapter focuses on the proposed efficient sampling method. Chapter 4 will present proposed contributions in multiple contingencies analysis and decision making process using decision tree.

2.4.1 Stage I - Identification of boundary region

Consider the sampling space to be an N-dimensional hypercube, where N is the number of selected operating parameters to be used in the study (loads, production levels, etc.). Stage I divides the hypercube into M smaller hybercubes. The situation for the simplest case, N=2, with M=20, is illustrated in Fig. 2.3.

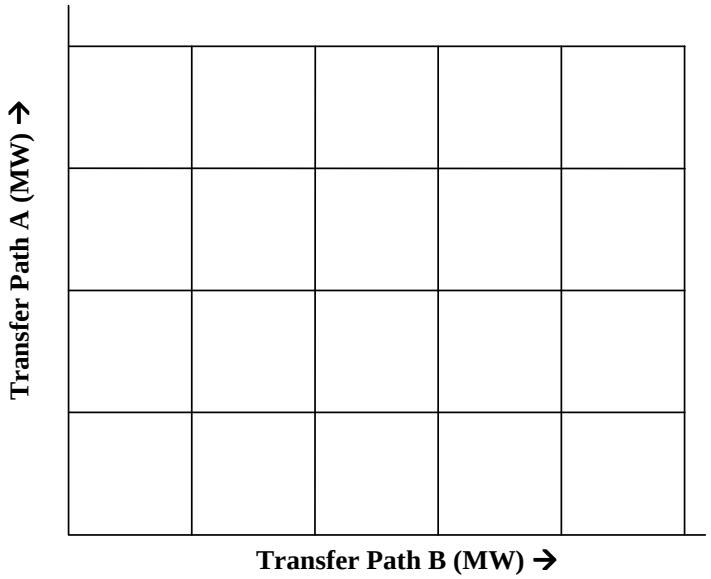


Fig. 2.3 Illustration of stratified sampling

Stage I selects the center point of each of the M smaller hypercubes and performs an assessment to identify post-contingency performance for each point. In other words, a first set of simulations is launched on a limited number M of network situations among all the possible ones at a coarse resolution. A typical result of such a sampling is shown in Fig. 2.4, where the enclosure contains all hypercubes that neighbor a hypercube of the opposite performance level, forming a first estimation of the boundary region.

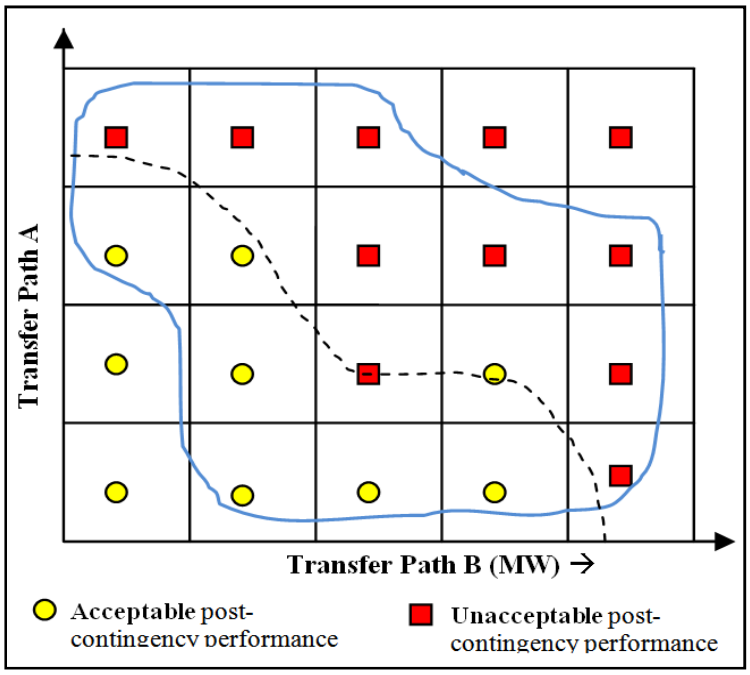


Fig. 2.4 Illustration of stage I

2.4.2 Stage II – Sampling

The standard Monte Carlo sampling approach draws values for each parameter in proportion to the assigned distribution. Given the previous knowledge of the boundary region from Stage I, biasing the sampling process towards the boundary region using importance sampling method maximizes the information content.

2.4.2.1 Importance Sampling Variance Reduction

In both adequacy studies and risk-based security planning studies, the quantity of interest is probability of unacceptable performance, i.e., *P(Y ~ unacceptable events)* [9].

$$P(Y < t) = \int_{-\infty}^{t} f(y)\,dy \tag{2.2}$$

where, $y=t$ denotes the threshold performance level such that $y < t$ is unacceptable performance. The indicator function *I(y)* denoting region of interest *h(y)* is defined as,

$$h(y) = I(Y < t) = \begin{cases} 1 \text{ if } Y < t \\ 0 \text{ if } Y \geq t \end{cases} \tag{2.3}$$

and hence,

$$P(Y < t) = \int_{-\infty}^{\infty} h(y) f(y) dy = E(h(Y)) = \sum_{i=1}^{n} h(y_i) \tag{2.4}$$

The above expectation function gives crude Monte Carlo estimation [65], where *yi* are Monte Carlo samples taken from the distribution *f(y)*, the post-contingency performance index probability distribution. This estimation has a variance associated with it, as the quantity $h(y_i)$ varies with y_i. Importance sampling attempts to reduce the variance of the crude Monte Carlo estimator by changing the distribution from which the actual sampling is carried out. Suppose it is possible to find a distribution *g*(y) such that $g(y) \alpha\ h(y) f(y)$, then the variance of estimation can be reduced by reformulating the expectation function as,

$$P(Y < t) = \int_{-\infty}^{\infty} h(y) f(y) \frac{g(y)}{g(y)} dy = E\left(\frac{h(Y) f(Y)}{g(Y)} \right) = \sum_{i=1}^{n} \frac{h(y_i) f(y_i)}{g(y_i)} \tag{2.5}$$

where y_i are Monte Carlo samples drawn from the distribution *g(y)*, and this ensures the quantity $\left[h(y_i) f(y_i) \big/ g(y_i) \right]$ is almost constant with y_i.

By choosing the sampling distribution *g(y)* this way, the probability mass is redistributed according to the relative importance of *y* as measured by the function |*h*(y)| *f*(y) [61].

2.4.2.2 Proposed Efficient Sample Generation

The property of importance sampling to bias the sampling using an importance function *g(y)* towards an area of interest, as discussed above is used to generate influential operating conditions from operational state space, *X* in our method. The joint probability distribution of the operational parameter space *f(x)* can be obtained from historical data [66].

Once we have *a-priori* information about *f(x)*, stage-I operation provides the region in *X* through which the boundary most likely occurs and therefore identifies approximately the *x*-space in which we want to bias the sample generation. The region of interest for sampling in terms of indicator function is,

$$h(X) = I(X \in S) = \begin{cases} 1 \text{ if } Y(X) \in S \\ 0 \text{ if } Y(X) \notin S \end{cases} \tag{2.6}$$

where S is the boundary region. For instance, in a univariate case, we can define it as $S=\{x: x_1 \leq x \leq x2\}$, as shown in Fig. 2.5.

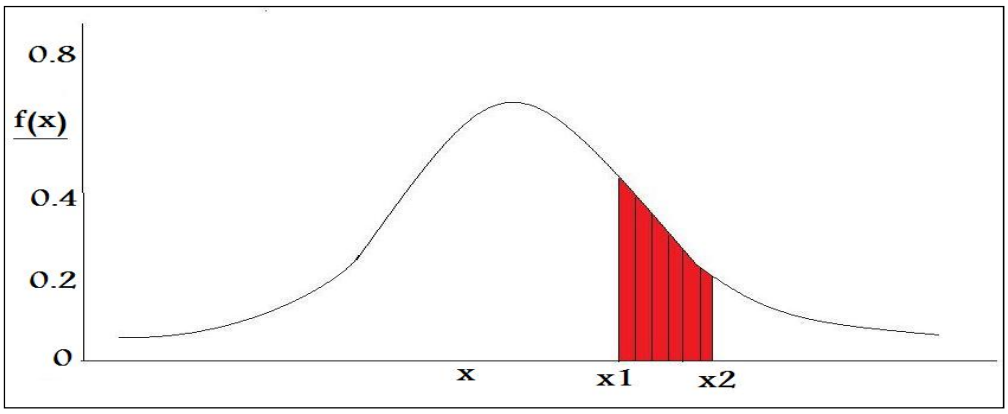


Fig. 2.5 Boundary region in operating parameter distribution f(x)

The importance function or the sampling distribution *g(x)* can be constructed proportional to |*h(x)*| *f(x)*, i.e., focusing on the region S of *f(x)*. In general, the importance sampling density can be expressed as,

$$g(x) = p * f_1(x) * \mathrm{I}(x \in S) + (1-p) * f_2(x) * \mathrm{I}(x \notin S) \tag{2.7}$$

where *p* controls the biasing satisfying the probability condition $p \leq 1$, $f_1(x)$ is the probability density function of the boundary region, and $f_2(x)$ is the probability distribution function of the region outside boundary.

We adopt a composition algorithm to generate samples from this distribution [67, 68]. If we set *p*=0.75, then 75% of the points can be expected from region S. This kind of upward scaling in boundary region probability distribution by the importance function *g(x)* is depicted by Fig 2.6.

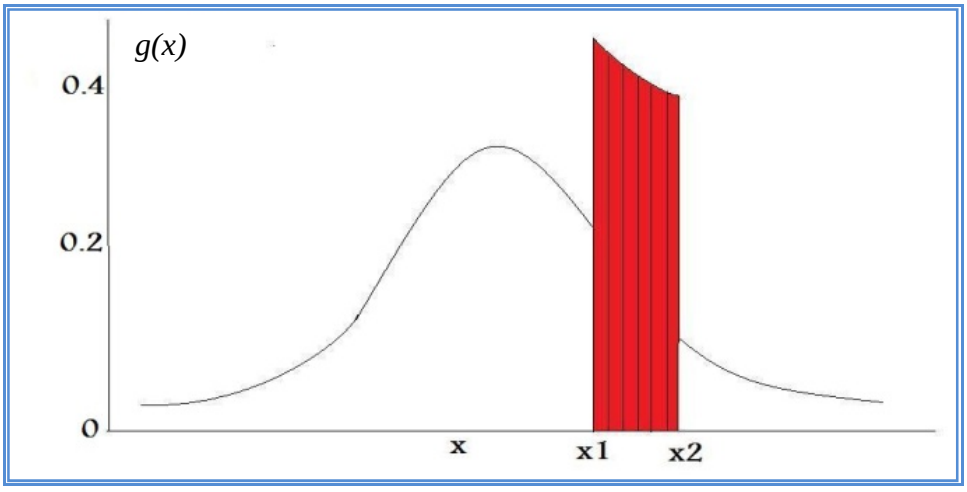


Fig. 2.6 Generic importance sampling distribution function *g(x)*

Hence *p* serves as sliding parameters that control the extent of biasing, i.e., sliding between a completely operational study (*p=1*, requiring most influential points for rule

making) to investment planning study (*p=0,* requiring a wide range of operating conditions). The optimal importance sampling density *g(x)* for our operational study is when *p*=1, i.e., full bias towards the boundary region, expressed as the original state space probability distribution conditional on the boundary region,

$$g(x) = f(x \mid x \in S) = \frac{1}{a} f(x) \tag{2.8}$$

$$a = \int_S f(x)dx \tag{2.9}$$

Since the scaling factor '*a*' is a probability and therefore, must obey $0 \leq a \leq 1$, equation (2.9) represents an upwards scaling. i.e., the probability distribution is altered such that more samples are from the region of interest.

2.5 Numerical Results

2.5.1 System Description

The proposed sampling approach is applied for a decision tree based security assessment study for deriving operating rules against voltage stability issues on SEO region (*Système Eléctrique Ouest*, West France, Brittany), a voltage security-limited region of the French EHV system containing 5331 buses with 432 generators supplying 83782 MW.

Figure 2.7 shows 400 KV network of the French system, where it can be seen that the Brittany region (in grey) is pretty weakly interconnected. During winter periods, when demand peaks, the system gets close to voltage collapse limits. Moreover the local production capabilities being far lower than the local consumption, it puts the EHV grid under pressure as the needed power comes from remote location, eventually leading to cascading phenomenon at the sub voltage levels. The red star indicates busbar fault at 225

KV Cordemais bus, which is the most credible contingency in the Brittany region during winter period.

Fig. 2.7 French 400 KV network with SEO and Brittany highlighted

So in order to avoid the risk of collapse situations under such contingency events, the operator may have to resort to expensive preventive measures such as starting up close yet expensive production units. It is therefore very important to assess the risks of a network situation correctly considering uncertainties in operating conditions and obtain operating rules built off-line with decision trees, that aid to take right decision at right time.

2.5.2 Study Specifications

Data preparation: The historical database of French EHV power grid system for the study is extracted from records made every 15 seconds on the network by SCADA, as shown by Fig. 2.8. The data for each month of the year is stored in many text files containing respecting columns of data:

- *Time data* i.e. the year, the month, the day, the hour, the minute, the day of the week of the recording;
- *Node data* i.e. voltage, voltage level, active and reactive consumption and production per node; and
- *Branch data* indicating the origin and the end nodes, their voltage level, if they are connected or not and the active and reactive transit considered at both extremities.

Figure 2.9 shows the 2007 annual load data in SEO region of French grid extracted from the historical database. The load starts to increase much at the end of October, as the winter comes closer, and decreases in February. The heavily loaded period is the winter, during December, January, and February months. A lot of loads were shed in the month of January under stressful conditions motivated by economic and reliability considerations for system operation, which explains the dip in the load during that month.

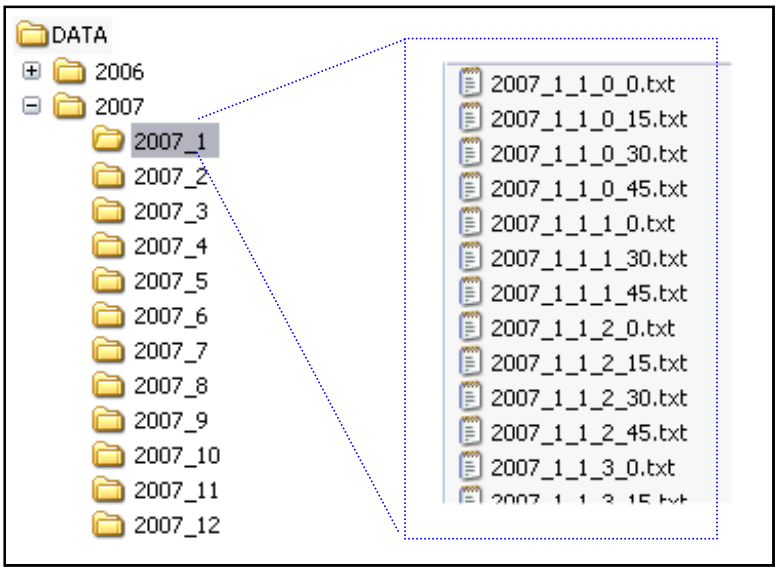


Fig. 2.8 French EHV historical data from SCADA

The loading pattern over the year changes depending upon various factors such as, if it is winter or summer, week or week-end, day or night, peak-hours or off peak hours etc. Typically, the load is heavier during the daytime of weekdays in winter, as shown by the statistics in Table 2.1. There are two peak-hours during a day in winter, i.e., in the morning around 8/9 am and the evening around 7.30/8 pm; and there is a secondary peak hour around 10/10.30 pm, as shown by Fig. 2.10 where a typical behaviour of the load over a typical winter day (7th February 2007) is depicted.

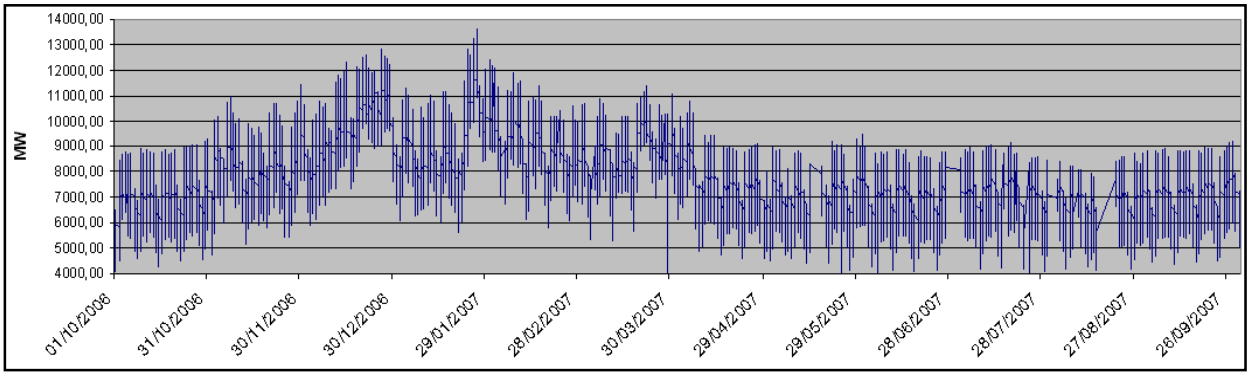


Fig. 2.9 2007 annual SEO load

Table 2.1 2007 historical load data statistics

	Mean	***Median***	***Max***
Full year	7729	7640	13607
Summer (June to Sept.)	6609	6600	9182
Winter (October to march)	8585	8539	13607
Winter (December to Feb.)	9290	9307	13607
Winter (December to Feb.) – Week days	9758	9823	13607
Winter (December to Feb.) - Week 8hr to 22hr	**10350**	**10284**	**13607**

Therefore, these heavily loaded periods are the most constraining in terms of voltage, and the study focuses on them for generating samples of operating conditions in the voltage stability study. Therefore, MCS is not performed on the entire year distribution, but only on those relevant periods of year depending on the type of stability problem under consideration.

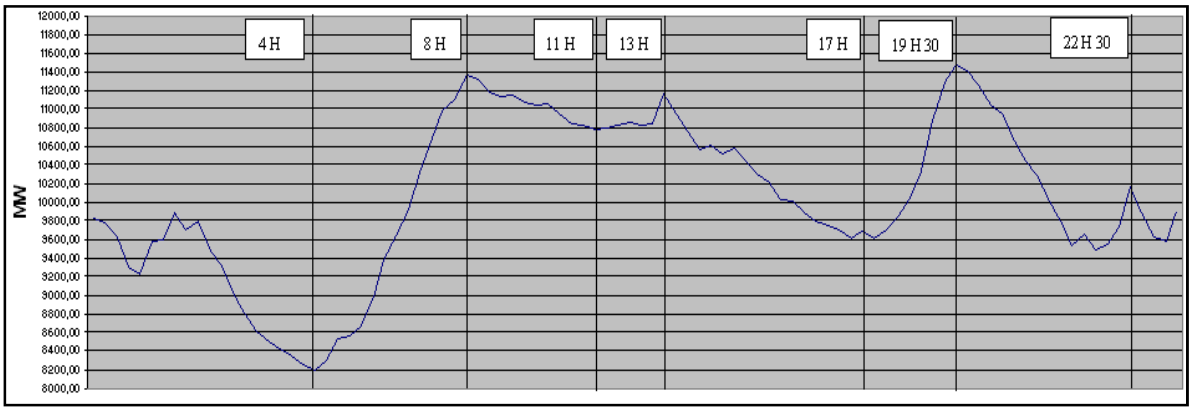


Fig. 2.10 Load behavior on February 7, 2007 – A typical winter day

Sampling: The pre-contingency operating conditions are generated from a base case, by considering random changes of key parameters. The basecase of SEO network considered corresponds to 2006/2007 winter with 13500 MW baseload. The most constraining contingency is the Cordemais busbar fault in the Brittany area that leads to trip nearby generation units. Under extreme conditions, this fault may lead to a Brittany voltage

collapse. The parameters that we play on to generate basecases are total SEO load, SVC unavailability and generator group unavailability in Brittany area. The unavailability of main production units, which includes nuclear groups in Civaux, Blayais, St-Laurent, Flamanville, and Chinon are sampled such that each of these 5 unavailabilities are represented in $1/6^{th}$ of the total basecases. There are 2 SVCs in the Brittany region i.e., at Plaine-Haute and Poteau-Rouge, and their unavailabilities are sampled such that 25% of the cases have them both, 25% do not have them both and 50% have only one of them. The total Brittany load, continuous parameter, is sampled using our proposed efficient sampling method. The load sampling is done keeping power factor constant. All the load samples are systematically combined with SVC and generator group unavailabilities respecting their respective sampling laws to form various basecases.

Contingency analysis and database generation: For each basecase, an optimal power flow is performed, minimizing the production cost under voltage, current, flow constraints in N. Abnormal/unrealistic cases that results in MW shedding or MVar addition to achieve convergence or do not converge are thrown off. Then consequences of busbar fault event are studied with a quasi steady state simulation (QSSS) tool, where the simulation is run for 1500s with 10s step size, and the contingency is applied at 900s. Scenarios are characterized as unacceptable if any of SEO EHV bus voltage falls below 0.8 p.u or the simulation does not converge. Then a learning dataset is formed using pre-contingency attributes of every scenario (sampled at 890s of QSSS) that drives voltage stability phenomenon, such as voltages, active/reactive power flows, productions etc, and their respective classifications. Then security rules are produced from decision tree to detect a probable voltage collapse

situation contingent upon the severe event. An independent test set is used to validate the tree.

The software tools used in the study are:

1. ASSESS [69] - Special platform for statistical and probabilistic analyses of power networks, that has the capability to generate many scenarios randomly or systematically to model system uncertainties

2. TROPIC [69] - Optimal Power Flow tool, embedded with ASSESS, to create initial base cases

3. ASTRE [69] - Simulating slow dynamic phenomena (QSSS), embedded with ASSESS

4.SAS - Statistical analysis and database processing

5. ORANGE [32], WEKA [33] - Decision tree tools

2.5.3 Efficient Sampling of Load Parameter

As mentioned in the section 2.5.2, the variable part of the system load, a continuous parameter that will accommodate the various uncertainties in the operating conditions was sampled according to the proposed efficient sampling method. The load is homothetically distributed among all the individual loads, i.e., a constant stress direction.

In order to find the boundary region in the load state space, a stratified sampling (100 MW interval) of the load was done, many variants were formed by systematically combining with discrete variables, i.e., SVC and generator unavailability. Contingency analysis was performed for every variant and each scenario is classified as acceptable and unacceptable. Figure 2.11 shows the characterization of boundary region in the load state space with respect to post-contingency performance. The boundary region capturing the variability of

performance measure is the defined by the range of values between 11860 MW and 12600 MW.

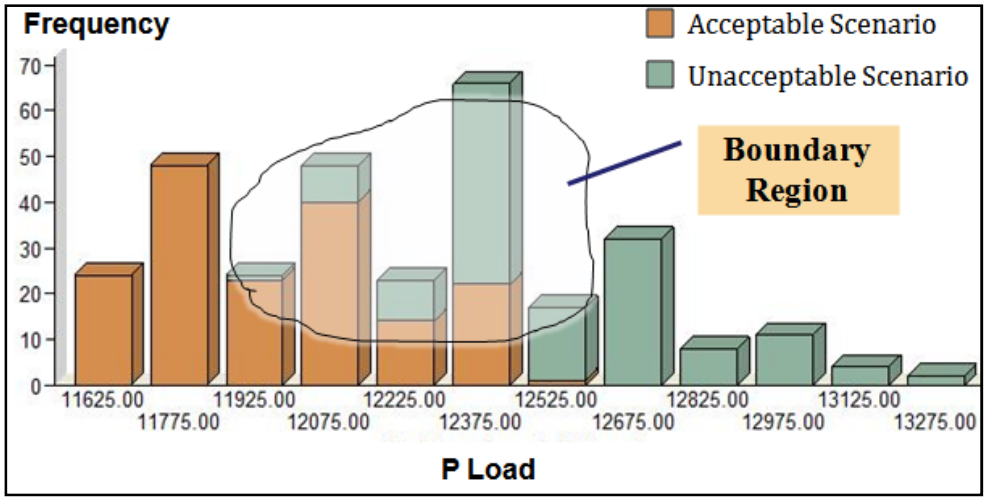


Fig. 2.11 Stratified sampling defining boundary region

The variable part of the system load, a univariate variable, follows a normal distribution *N(9883.6, 979583)*, according to 2006-07 historical data of peak hours (weekdays 8hr to 22hr) during winter period, as shown in Fig 2.12. Figure 2.12 also shows the probability distribution of the boundary region identified by stage-I. Importance sampling is performed on the probability distribution of the load with $p = 1$ in equation (2.7), to bias sampling towards the boundary region.

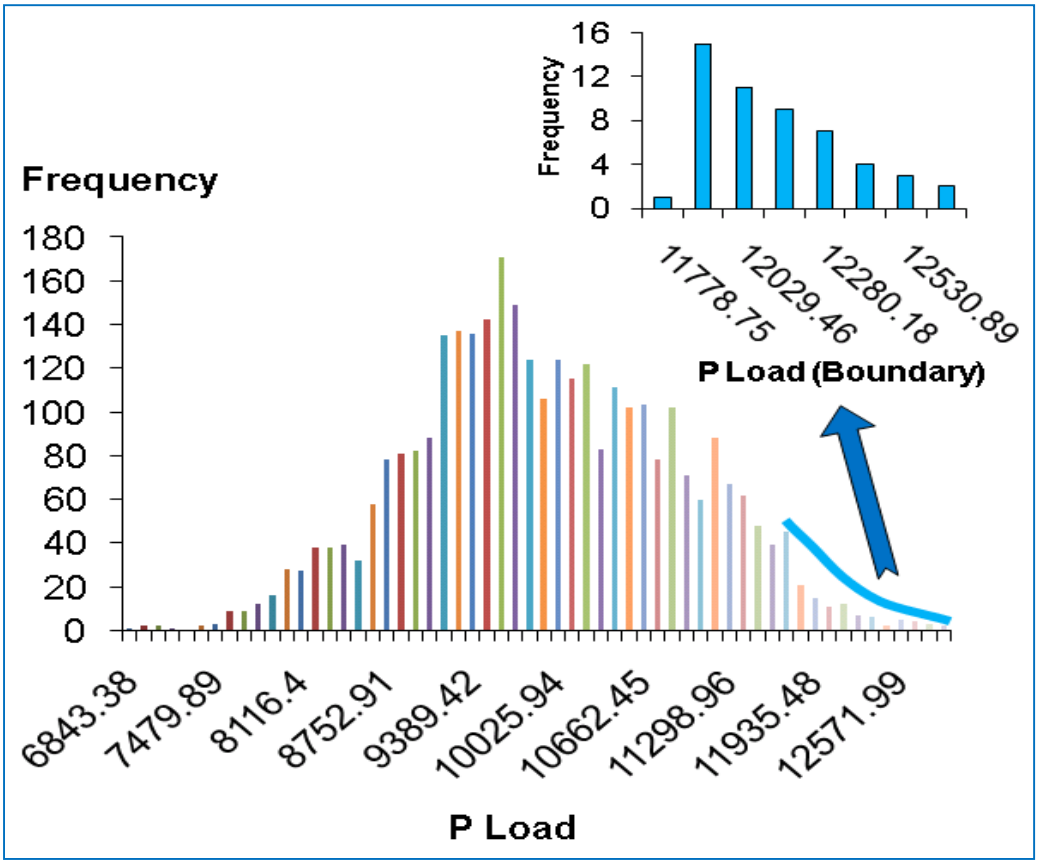


Fig. 2.12 Probability distribution of variable part of the system load

2.5.4 Results

There are many system attributes that can be included in training dataset as potential rule attributes. Some of them that are influential for a voltage stability study include 400 KV node voltages, active and reactive power reserves of the production groups in SEO, active and reactive power flows in tie lines of SEO, net inter-area transactions, etc.

Table 2.2 shows the effectiveness of various attribute sets in terms of classification accuracy and error rates. Accuracy can be defined as the percentage of points correctly classified, false alarm rate can be defined as the ratio of total misclassified unacceptable instances among all unacceptable classifications, and risk rate is defined as the ratio of total misclassified acceptable instances among all acceptable classifications. Attribute set

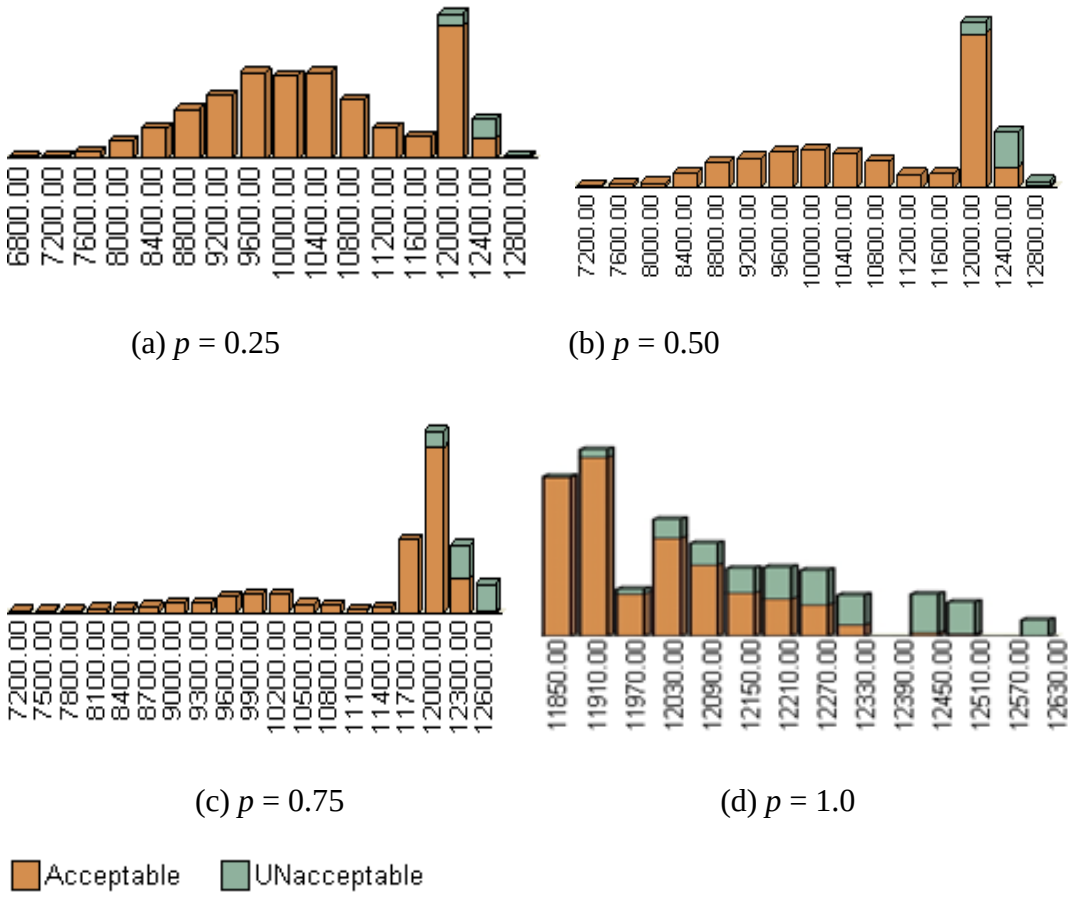
"Voltage" contains 46 400 KV node voltages, "P reserve" contains 10 generator group's and total SEO real power reserve, "Q flow" contains various attributes such as 12 400 KV tie line reactive flows from SEO region to other regions, 4 interarea 400 KV reactive transfers, and net reactive power export; and "Q reserve" contains 10 generator group's and total SEO (including SVCs) reactive power reserve.

The training database obtained by sampling from the boundary region contains 940 operating conditions. The test set includes 459 instances unseen by training set that covers a wide range of operating conditions, with some also falling within the boundary region, for it is very important to obtain decision rules that classify conditions near the threshold correctly. From Table 2.2, we can see that "Q reserve" is a good attribute with lowest risk among high accuracy attributes. This conclusion meets local operators' experiences that Q reserves give warning prior any voltage drop.

Table 2.2 Attribute set performance comparison

Attribute set	***Accuracy (%)***	***False Alarm***	***Risk***
Voltage	98.4	0.013	0.023
P reserve	90.1	0.059	0.201
Q flow	97.34	0.025	0.03
Q reserve	99.04	0.006	0.019
Voltage + Q flow	98.83	0.01	0.015
Voltage + Q reserve	99.04	0.004	0.023
Voltage + Q reserve + Q flow	99.04	0.004	0.023

Figures 2.13 (a), (b), (c), (d) show the total SEO load probability distribution from sampled operating conditions as the sliding factor *p* increases from base value in *f(x)* to 1.



(a) p = 0.25 (b) p = 0.50

(c) p = 0.75 (d) p = 1.0

Fig. 2.13 Effect of p on sampled total SEO load probability distribution

This study was performed to investigate the influence of the sliding factor p on rule performance. Table 2.3 shows the results, when validated using the *test dataset* mentioned earlier. A slight bias in the test set distribution towards security boundary region is to validate the operational rule's classification performance against critical scenarios and also to show the significance of generating high information contained training database. Nevertheless the test set is still independent due to the fact that the testing samples are generated randomly and the instances are unseen by the training set.

In Table 2.3, we can see that the training database biasing towards boundary region increases as sliding factor k_1 increases from default value of about 15% (in the original

distribution) to 100%, as observed from the fact that the representation of unacceptable scenarios (Un) relative to acceptable scenarios (A) increases in the database of same size. Consequently the value of entropy, computed according to equation (1) measuring the information content in the database, also increases as the samples generated from boundary region increases.

Table 2.3 Performance comparisons between sampling bias

Bias, p (%)	***A:Un***	***Entropy***	***Accuracy (%)***	***False Alarm***	***Risk***
15 (base)	889:51	0.3042	83.19	0.033	0.527
25	825:115	0.5361	95.21	0.028	0.087
50	781:159	0.6558	96.17	0.027	0.068
75	738:202	0.7507	97.55	0.016	0.045
100	676:264	0.8566	99.04	0.004	0.023

Figure 2.14 shows the increase in rule accuracy, and Fig. 2.15 shows the decrease in false alarm and risk rate, with increase in bias towards boundary, indicating that the training set generated within boundary can classify well wide-range of operating conditions. This is very beneficial for an operational planning study. Similarly, by suitably adjusting k_1, we can draw operating conditions that cover a wide range in parameter state space suitable for investment planning studies.

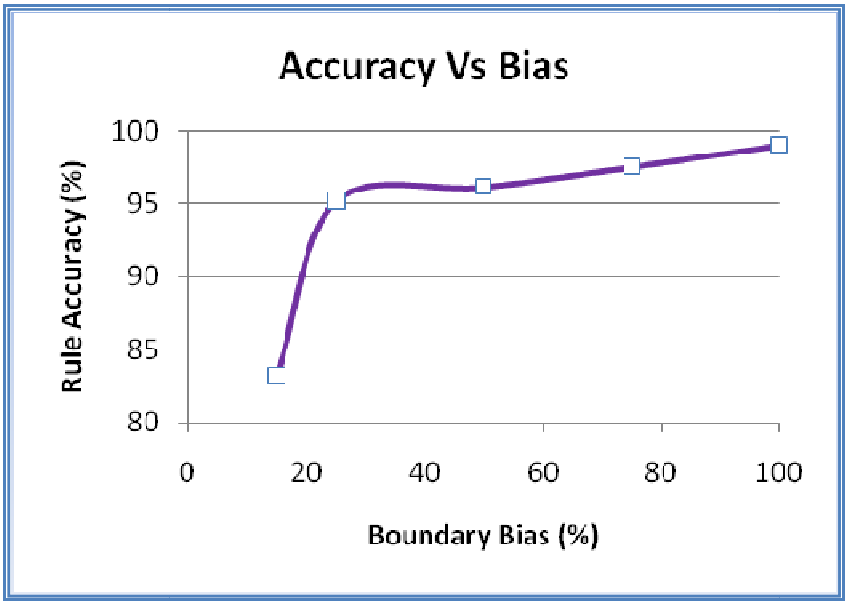


Fig. 2.14 Rule accuracy vs. sampling bias towards boundary

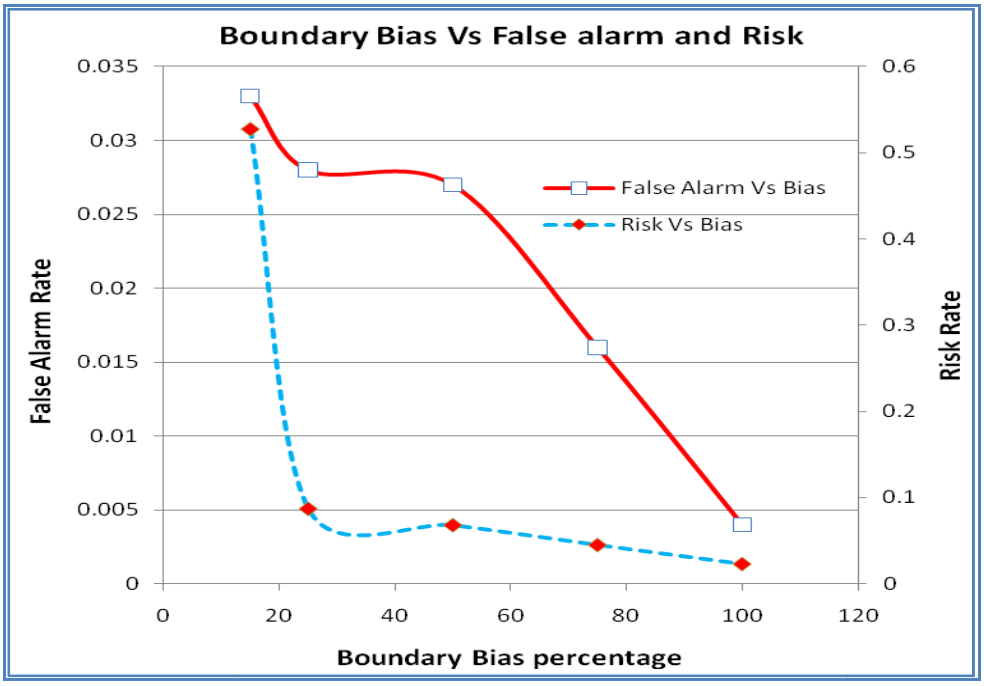


Fig. 2.15 Error rates vs. sampling bias towards boundary

Figure 2.16 shows the plot between classification accuracy and entropy as the bias factor k_1 increases from base value to 100%, for a given database size of 940, i.e., for a constant computing requirement. It can be seen that the classification accuracy increases as the

training database entropy increases. This indicates that for a given computation the database that exclusively captures the variability of performance measure across the boundary region performs well.

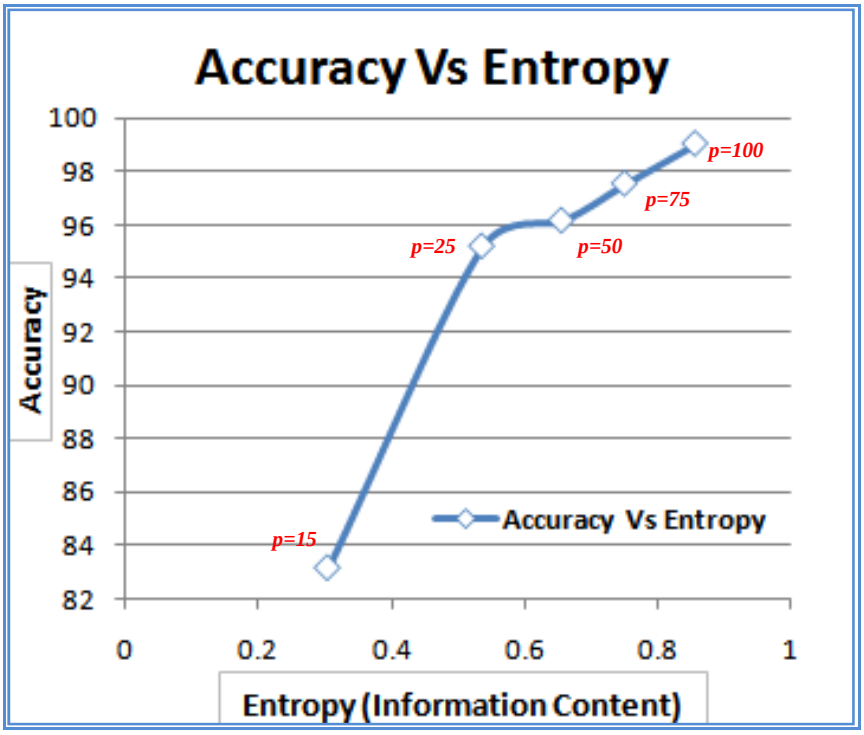


Fig. 2.16 Accuracy vs. database entropy, for a given computation

Table 2.4 shows the result of another study comparing three different sampling approaches, namely, sampling from the entire state space according to its probability distribution, uniform sampling of boundary region, and importance sampling of boundary region. It can be seen that the accuracy is more and the error rates are less for importance sampling, even with decreased computation, as depicted by Fig. 2.17. Figure 2.17 also shows that by increasing computation deliberately, higher accuracy can be obtained with importance sampling strategy.

Table 2.4 Performance comparisons between different sampling strategies

Sampling	*Size*	*Accuracy (%)*	*False Alarm*	*Risk*
1. Entire Space	940	83.19	0.028	0.527
2. Boundary Uniform	800	92.35	0.11	0.043
3. Boundary IS-I	470	94.89	0.028	0.11
4. Boundary IS-II	752	96.81	0.013	0.08
5. Boundary IS-III	940	99.04	0.004	0.023

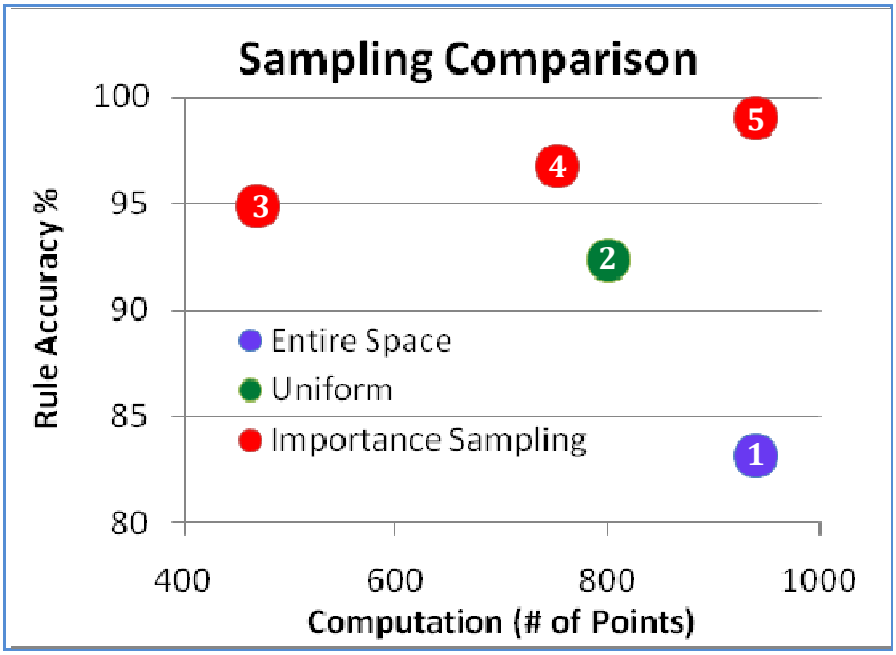


Fig. 2.17 Comparison between sampling strategies

The above results show the effectiveness of importance sampling based strategy to generate efficient training set for decision tree based learning studies. It was observed that with lesser computation more information content can be generated, and consequently improvement of operating rule's performance is possible. The developed training database generation method can be applied for other data mining techniques as shown in Table 2.5, and also against other power system security problems.

Table 2.5 Importance sampling for various data mining techniques

Bias factor, p	Naïve Bayes	SVM[1]	IB5[1]	DT[1]
0.25	75.79	96.48	94.07	95.21
0.5	78.71	98.28	95.45	96.17
0.75	83.43	98.71	97.16	97.55
1	92.78	99.65	97.68	99.04

Typically, a rule is desired to be simple and efficient enough to separate unacceptable situations from acceptable ones, such that it leads to no risks and minimizes the false alarms. The risk corresponding to importance sampling method (No.5) shown in Table 2.4 with 940 samples is 0.023%. One way to reduce risks is to use a cost-sensitive classification, i.e., specifying a cost for misclassification. By making the cost of risk twice the cost of false alarm, the risk percentage is reduced to 0.011%, while false alarm slightly increases to 0.01% from 0.004%. The cost of misclassification reduces by 2 units, under the assumptions of cost.

Another way to reduce risk is to have a feedback loop from the rule validation stage to sample generation stage, which gives appropriate information to increase the representation of expensive misclassified conditions in the database, so that the decision tree is able to classify them properly. In real time application, the misclassified or strange (i.e., in comparison with the historical loading conditions) operating conditions can be flagged and then used to update the decision rules by updating the training database with the flagged instances.

[1] SVM - support vector machine; IB*5* - nearest *5*-neighbour instances based learning; DT - decision tree

2.6 CONCLUSIONS

The proposed efficient sampling method based on importance sampling idea is one of the first to be used in power systems for making decision tree based learning methods effective. The thrust of the proposed sampling procedure is to re-orient the sampling process using importance sampling to focus more heavily on points for which post-contingency performance is close to the threshold forming the boundary region that contains operating conditions influential for rule formation. The primary goal is to increase the information content in the learning database while reducing the computing requirements, and consequently obtain operational rules that are more accurate for usage in real-time situations. The results show that the generated training database enhances rules' accuracy giving less error rates when compared with traditional sampling approaches.

CHAPTER 3 EFFICIENT PROCESSING OF SYSTEM SCENARIOS IN MULTIVARIATE NON-PARAMETRIC OPERATING PARAMETER DISTRIBUTION

3.1 INTRODUCTION

Decision tree based planning tools provide operators with the most important system attributes that guide them in deciding as to what situation requires operator action. Chapter 2 focused on the key aspect of this approach, namely devising an efficient Monte Carlo sampling approach to capture high information content and reduce computational cost in the database generation step. The developed efficient sampling process was also illustrated on French EHV network. This chapter focuses on the data processing (preparation) stage prior to the MCS stage, and the techniques to achieve the proposed efficient Monte Carlo sampling approach are appropriately constructed.

3.2 MOTIVATION AND PROPOSAL

In chapter 2, the global load was distributed homothetically (i.e., proportion of individual loads to global load same as basecase) along the most probable stress direction. This is typically done in various studies, where samples of representative basecases are drawn for various loading conditions, i.e., peak, mid, low etc., assuming a particular load stress pattern. Some of the motivations for such assumption are:

1. The assumed stress direction is the most likely one as indicated by the historical data.
2. To reduce the computational burden.

The sampling procedure becomes computationally very burdensome for a very large dimensional sampling state space, if the individual load's mutual correlation information is

taken into account for accommodating multiple stress directions. So, in order to provide a more reasonable sampling space which would reduce the computation, a very strong assumption is made that all loads vary in proportion to the total, so that the load at any bus *i* maintains a constant percentage of total load as total load changes, i.e., $P_{Li} = \frac{P_{Li0}}{P_{T0}} P_T$, where P_{Li0} and P_{T0} are the bus *i* load and total load, respectively, in the base case. In the language of voltage instability analysis, these assumptions amount to the definition of a particular *stress direction* through the space of possible load increases.

So the load uncertainty is addressed only in terms of a single variable: total load (P_T). This is illustrated in Fig. 3.1, where we consider a much simplified power system with only two load buses, and the mean value of P_T is the baseload of 1000 MW.

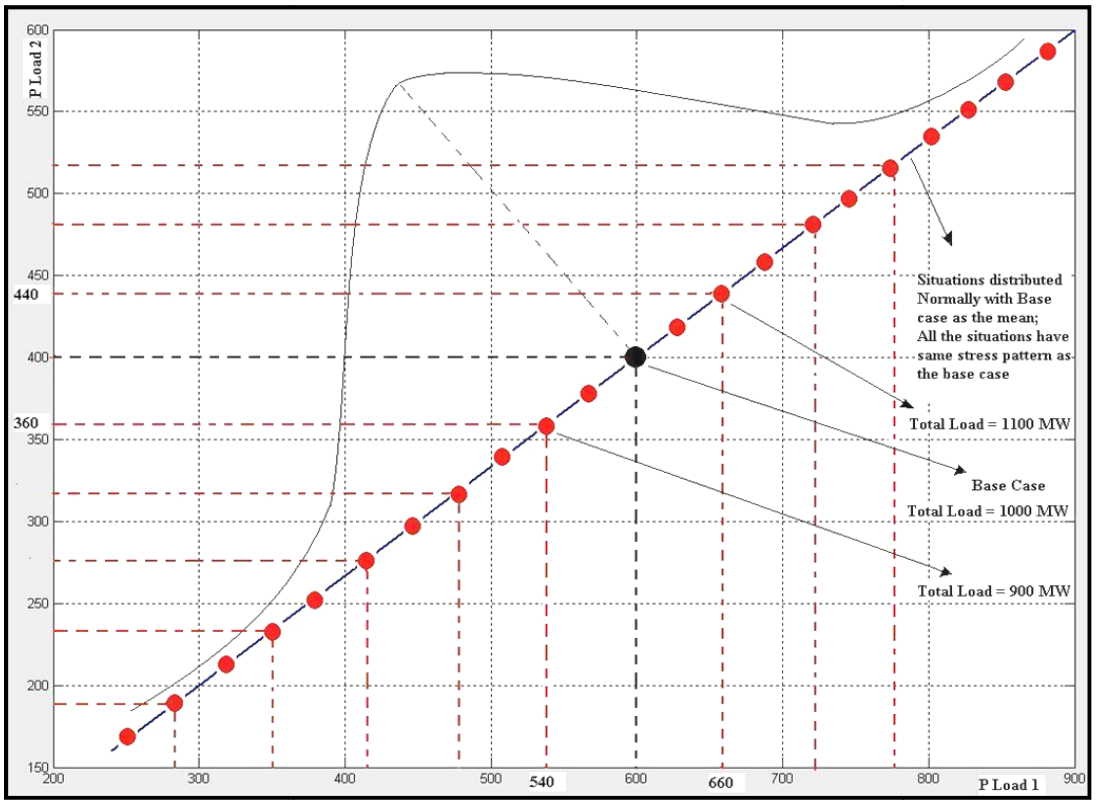


Fig. 3.1 Sample points of P_T in 2-dimensional parameter space with assumed stress direction

P_T is assumed to be distributed normally about its mean value, and the stress direction is defined by the assumed proportions of 60% and 40% for loads 1 and 2 respectively. So as the proposed efficient sampling approach was illustrated in chapter 2, the stratified sampling is performed only in the univariate space of total system load to identify the boundary region as shown in Fig. 3.2 and importance sampling is performed to bias the sampling towards this region.

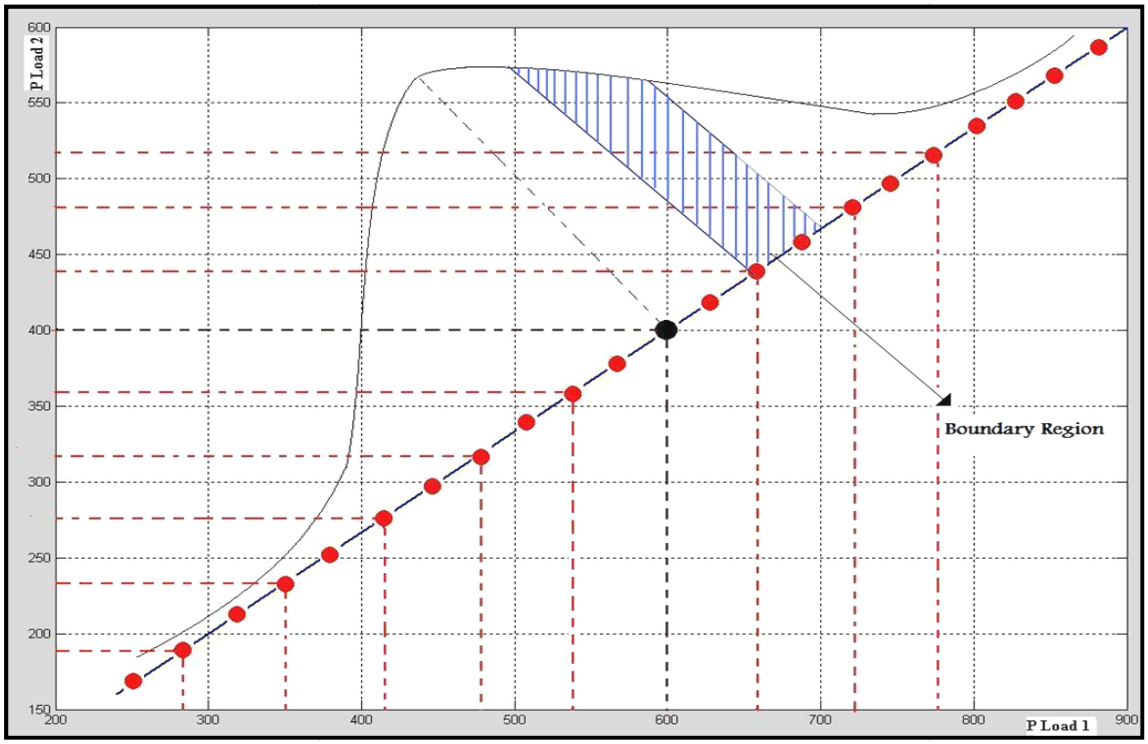


Fig. 3.2 Boundary identification within sample space of operating points shown in 2-D

However, in reality the individual loads may vary along multiple stress directions, and confining to the single stress direction may result in sampling too narrowly. Consider Fig. 3.3, which shows an operating parameter space for a three-load power system (and therefore it is a 3-dimensional figure). As discussed previously, sampling from a single stress direction (i.e., the expected stress direction) will result in a collinear set of points within the 3-D

figure, as shown by the line with red circles in Fig. 3.3. However, there may exist other operating points in the sample space, close to but not on the expected stress direction line, that are reasonably likely to occur compared to the points on the red line. For example, we may conceive of a region surrounding the expected stress direction line that contains points comprising a 0.95 probability space, i.e., the probability of occurrence of an operating condition outside that region is 0.05. Such a region is conceptualized in the three-dimensional picture of Fig. 3.3 as the "cylinder" confined by the two red dashed lines. The limits that define the boundary (between acceptable/unacceptable domains) would then become a surface cutting through this cylinder, as illustrated by the green surface in Fig. 3.3.

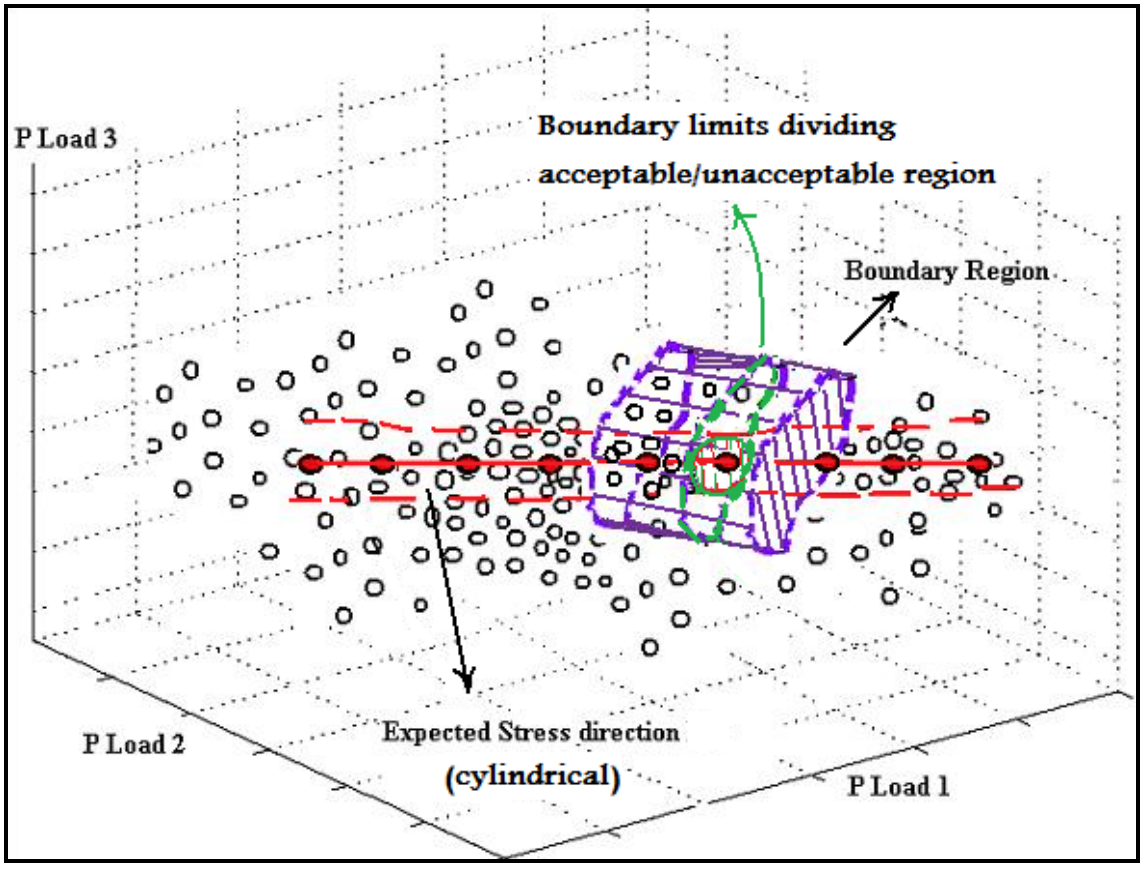


Fig. 3.3 Prospective boundary region in 3-D operating parameter sample space

Through the stratified sampling stage we would want to obtain the boundary region depicted by the 3D purple region in Fig. 3.3. Then the importance sampling could be applied to sample points within this boundary region, which would capture maximum information content including the relative likelihood of sample points.

The same concept can be illustrated as a 2-Dimensional example, depicted in Fig. 3.4 below. Even though the points seem to be following a single primary stress pattern, there are other sample points in the multivariate space that would be within a defined probability space. So it is important to consider the multivariate distribution of loads to capture the boundary region effectively, and capture high information content. Otherwise, single stress direction assumption will identify only some portion of boundary, and consequently the rules derived from such a database may face challenges when applied to realistic operating conditions, where we could expect loads to follow any stress pattern.

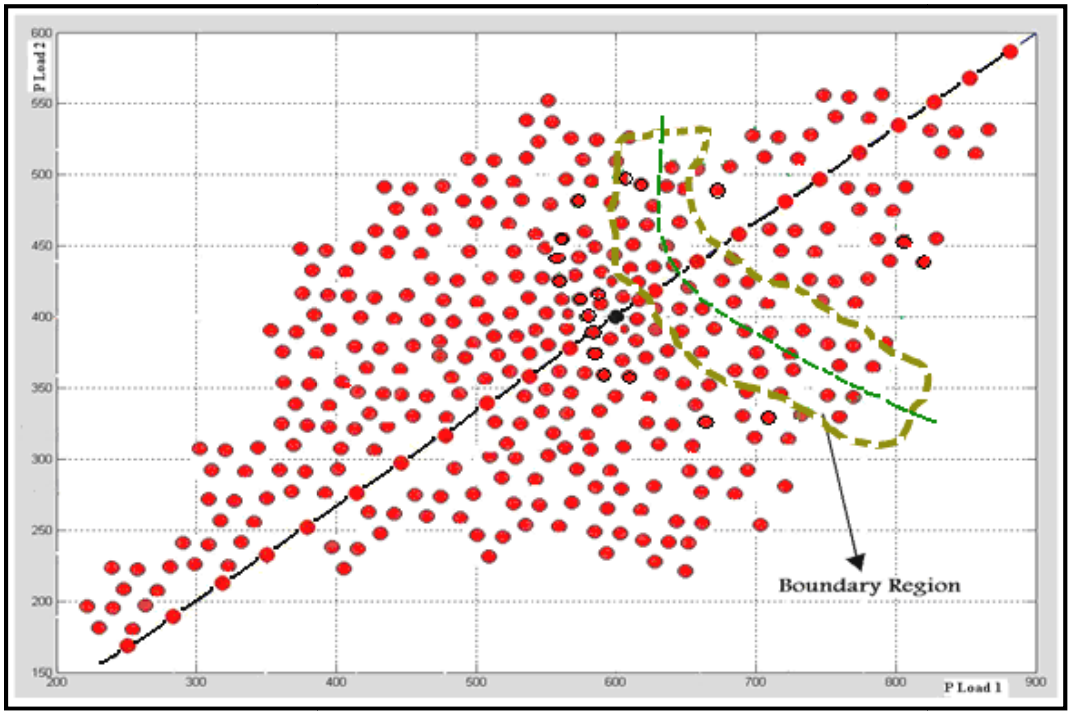


Fig. 3.4 Prospective boundary region in 2-D operating parameter sample space

Therefore, it is necessary to capture inter-load correlations from historical information while sampling from multivariate load distribution to create the training database, where such finer details will have crucial impact in a decision tree's ability to find rules suitable for realistic scenarios. While we can be assured of more information content, it is likely to increase computing requirements; especially for boundary identification stage using stratified sampling. Dobson et. al. [70] proposed a direct and iterative method to find the closest voltage collapse point with reduced computation in the hyperspace defined by loads. But the method's applicability to a specific distribution of loading conditions in the hyperspace was not shown, and doubts were also cast over its applicability to a large power system with dimension of the hyperspace going in 100s as we are dealing in this dissertation. In this chapter, we propose Monte Carlo simulation based method to find the stability boundary in a multivariate load state space at a highly reduced computational requirement. The reduction in computational cost is possible by the use of Latin hypercube sampling (LHS) of homothetic stress directions and linear sensitivities. The multivariate load state space for a given historical distribution is then quickly characterized, under various combinations of SVC and generator unavailability states. Then, we apply importance sampling to bias the sampling towards the identified boundary region.

In this study, we propose to model inter-load correlations in Monte Carlo simulation using copulas [71], unlike many studies that approximate the inter-load correlations using multivariate Normal distribution for computational purposes. Copulas are generated based on non-parametric historical load distribution, and it enables sampling realistic scenarios. The proposed method is envisioned to reduce the computational cost, while producing training

database with high information content that enables deriving operating rules with better knowledge of boundary limits, leading to higher classification accuracy, and economic rules.

The remaining parts of this chapter are organized as follows. Section 3.3 presents the technical approach, section 3.4 presents the application results of the proposed method in a voltage stability assessment for French power system, and section 3.5 concludes.

3.3 Technical Approach

The efficient sampling algorithm proposed consists of two stages, stage I to approximately identify the boundary region and stage II to bias the sampling towards the boundary region as shown in Fig. 3.5.

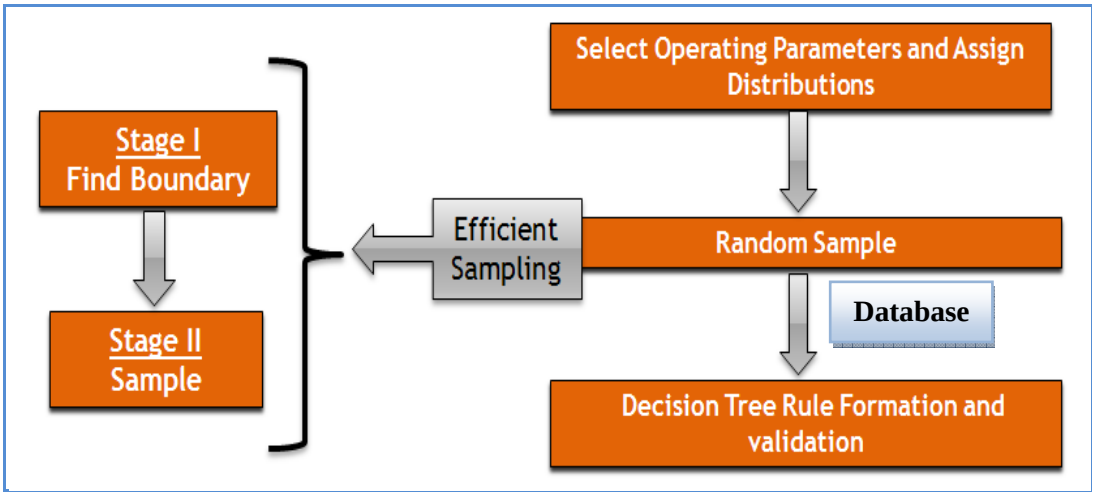


Fig. 3.5 Proposed efficient sampling algorithm

3.3.1 Stage I - Identification of Boundary Region

A straight forward way to perform state space characterization is to divide the N-dimensional hypercube, where N is the number of selected operating parameters, into M smaller hypercubes, select the center point of each of the M smaller hypercubes and perform an assessment to identify post-contingency performance (N^M contingency simulations), as

described in chapter 2. But for large N, there is a curse of dimensionality, resulting in very large computational cost. So this section develops a Latin Hypercube sampling method that uses linear sensitivity information to apply the developed efficient sampling approach in a computationally effective manner.

3.3.1.1 Fast Boundary Region Identification using Linear Sensitivity Information

For some performance measures, it is possible to use linear sensitivities to efficiently obtain improved approximation of the boundary between acceptable and unacceptable performance, as shown in Fig. 2.4 by the dotted line. This significantly reduces the computation burden in characterizing a multi-dimensional operational parameter state space. For voltage stability related problems, voltage stability margin (VSM) can be used as the performance measure and hence voltage stability margin sensitivities [72, 73, 74] with respect to operational parameters such as individual loads ($\partial VSM/\partial P_j$), generator availability, etc. can be used to identify the boundary.

Voltage Stability Margin: Voltage stability margin is defined as the amount of additional load in a specific pattern of load increase (also termed as stress direction) that would cause voltage instability as shown in Fig. 3.6. It is computed using the continuation power flow (CPF) method. Contingencies such as unexpected component outages (generator, transformer, transmission line etc.) in an electric power system often reduce the voltage stability margin [75, 76], and may cause the voltage stability margin to be negative (i.e. voltage instability) if they are severe. Figure 3.6 shows the voltage stability margin under different operating conditions.

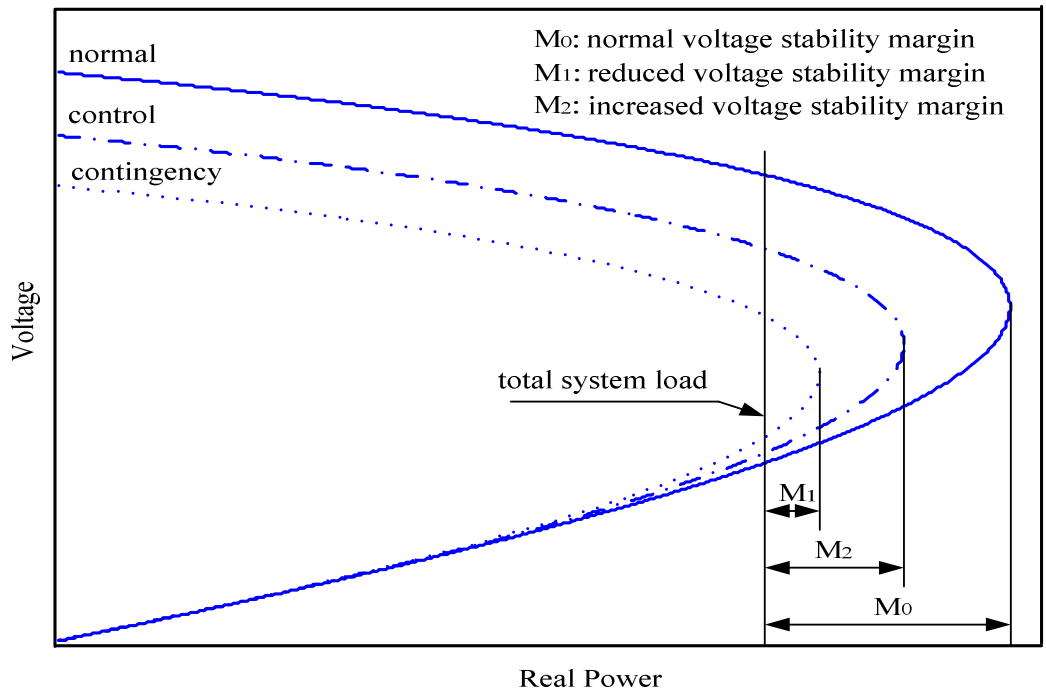


Fig. 3.6 Voltage stability margin under different conditions [77]

Voltage Stability Margin Sensitivity: The sensitivity of voltage stability margin refers to how much the stability margin changes for a small change in system parameters such as P and Q bus injections, regulated bus voltages, Bus shunt capacitance, Line series capacitance etc. It is computed as a by-product of the CPF computation to find the voltage collapse point, where the eigenvalues of the jacobian at the critical collapse point would give these linear sensitivities. Sensitivity computations have been typically used for two major purposes, contingency ranking and evaluating control action effectiveness [78].

Continuation Power Flow and sensitivity computation: Let the steady state of the power system satisfying a set of equations in the vector form be,

$$F(x, p, \lambda) = 0 \tag{3.1}$$

where, x is the vector of state variables, p is any parameter in the power system steady state equations such as demand and base generation or the susceptance of shunt capacitors or the

reactance of series capacitors, the state vector, and λ denotes the system load/generation level called the scalar bifurcation parameter. The system reaches a state of voltage collapse, when λ hits its maximum value (the nose point of the system PV curve), and the value of the bifurcation parameter is equal to λ^*. For this reason, the system equation at equilibrium state is parameterized by this bifurcation parameter λ as shown below.

$$P_{li} = (1 + K_{lpi}\lambda)P_{li0} \tag{3.2}$$

$$Q_{li} = (1 + K_{lqi}\lambda)Q_{li0} \tag{3.3}$$

$$P_{gj} = (1 + K_{gj}\lambda)P_{gj0} \tag{3.4}$$

where, P_{li0} and Q_{li0} are the initial loading conditions at the base case corresponding to λ=0. K_{lpi} and K_{lqi} are factors characterizing the load increase pattern (stress direction). P_{gj0} is the real power generation at bus j at the base case. K_{gj} represents the generator load pick-up factor.

When system parameters are changed, the total transfer capability will probably increase or decrease. Reference [79] explains margin sensitivity in the framework of DAE formulation,

$$\dot{x} = F(x, y, p) \tag{3.5}$$

$$0 = G(x, y, p) \tag{3.6}$$

where x are the state variables $x \in \overset{n}{R}$; y are the algebraic variables $y \in \overset{m}{R}$; p are the independent variables or parameters $p \in \overset{l}{R}$; f are the differential equations $f : \overset{n}{R}^* \overset{m}{R}^* \overset{l}{R} \to \overset{n}{R}$; and g are the algebraic equations $g : \overset{n}{R}^* \overset{m}{R}^* \overset{l}{R} \to \overset{m}{R}$.

$$\frac{\partial \lambda}{\partial P} = \frac{(w_F^T, w_G^T)\begin{pmatrix} F_P \\ G_P \end{pmatrix}}{(w_F^T, w_G^T)\begin{pmatrix} F_\lambda \\ G_\lambda \end{pmatrix}} \tag{3.7}$$

where w are the left eigenvectors of the Jacobian at the nose point.

Once $\partial\lambda/\partial P$ is computed, we will first get the bifurcation parameter estimation as

$$\Delta\lambda = \frac{\partial \lambda}{\partial P}\Delta P \tag{3.8}$$

For a power system model using ordinary algebraic equations, the bifurcation point sensitivity with respect to the control variable p_i evaluated at the saddle-node bifurcation point is

$$\frac{\partial \lambda^*}{\partial p_i} = -\frac{w^* F_{p_i}^*}{w^* F_\lambda^*} \tag{3.9}$$

where w is the left eigenvector corresponding to the zero eigenvalue of the system Jacobian F_x, F_λ is the derivative of F with respect to the bifurcation parameter λ and F_{p_i} is the derivative of F with respect to the control variable parameter p_i.

This margin sensitivity gives the first order partial derivative in the Taylor series expansion of λ as a nonlinear function of P, which describes the hypersurface $\sum$. The bifurcation parameter sensitivity will allow us to know, when some parameters are varied, how the system will move along the hypersurface $\sum$ in the vicinity of the current instability point denoted by λ_*.

The voltage stability margin can be expressed as [77]

$$M = \sum_{i=1}^{n} P_{li}^* - \sum_{i=1}^{n} P_{li0} = \lambda^* \sum_{i=1}^{n} K_{lpi} P_{li0} \tag{3.10}$$

The sensitivity of the voltage stability margin with respect to the control variable at location *i*, S_i, is

$$S_i = \frac{\partial M}{\partial p_i} = \frac{\partial \lambda^*}{\partial p_i} \sum_{i=1}^{n} K_{lpi} P_{li0} \tag{3.11}$$

The discussed concept is depicted in Fig. 3.7.

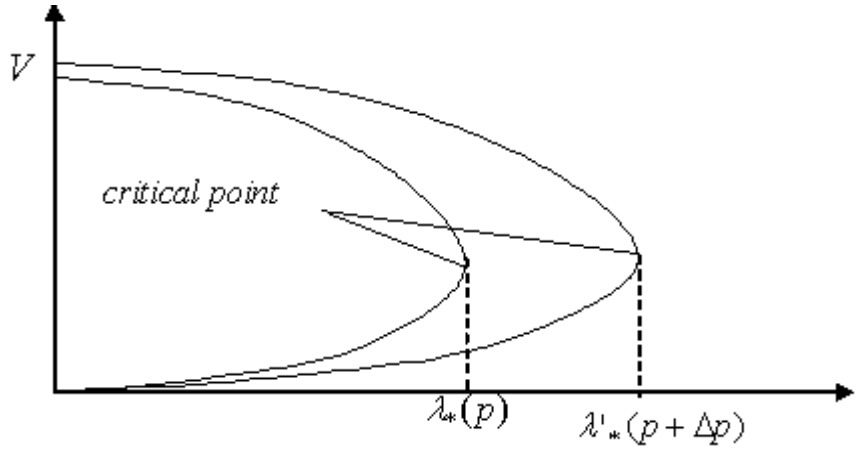


Fig. 3.7 Transfer margin change with the change of parameter, p [79]

The voltage stability margin and its sensitivity is computed using continuation power flow (CPF) method [80], as conventional power flow methods do not give any solution at the critical point due to singularity of power flow jacobian. In continuation method, the system equation at equilibrium state is parameterized by this bifurcation parameter λ, which is the scalar bifurcation parameter that parameterizes the load level. The system reaches a state of voltage collapse, when λ hits its maximum value (the critical point of the system PV curve as shown in Fig. 3.7), and the value of the bifurcation parameter is equal to λ_*, which gives the corresponding maximum loadability and hence the stability margin *M*. The bifurcation parameter sensitivity S^p with respect to control parameter *p* is obtained as a by-product of the continuation method.

3.3.1.2 Homothetic Stress Directions, Linear Sensitivities and Boundary Identification

The assumption of a stress direction is important to perform CPF study for identifying the voltage collapse point in that direction. The stress direction for performing CPF is defined by a particular combination of base load stress factors $P_i \Big/ \sum_{i=1}^{n} P_i$, $i=1,2\ldots n$ loads, as defined in section 3.2. Figure 3.8 shows the increase of total system load in a particular stress direction defined by the combination of three individual loads PL_1, PL_2 and PL_3.

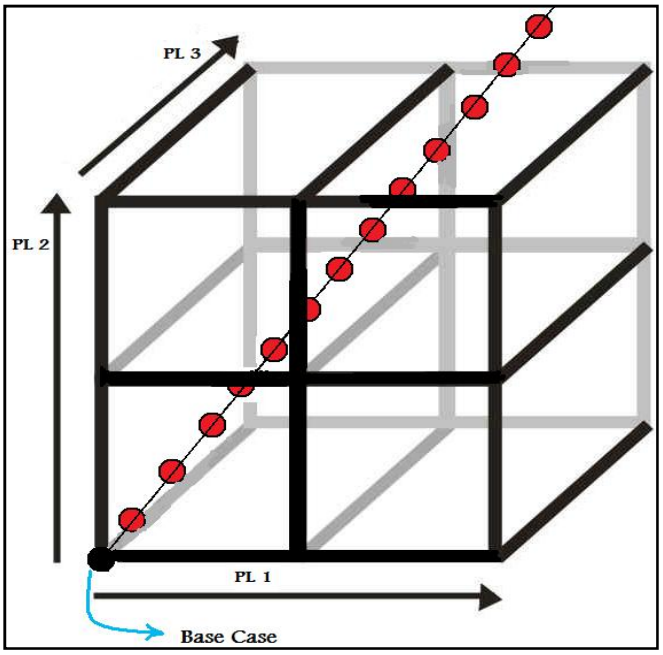


Fig. 3.8 Load increase in a particular stress direction

Such a distribution of stress due to increasing load is known as homothetic distribution of load (i.e., load repartition between the nodes same as the base case's intrinsic load factors). Figure 3.9 depicts this concept in two dimensional space defined by loads A and B. The line $Load_A+Load_B=C$ defines various basecases with different inter-node repartitions among

loads A and B for the same baseload C. These basecases define various homothetic stress directions in the state space, as shown by the various lines from the origin.

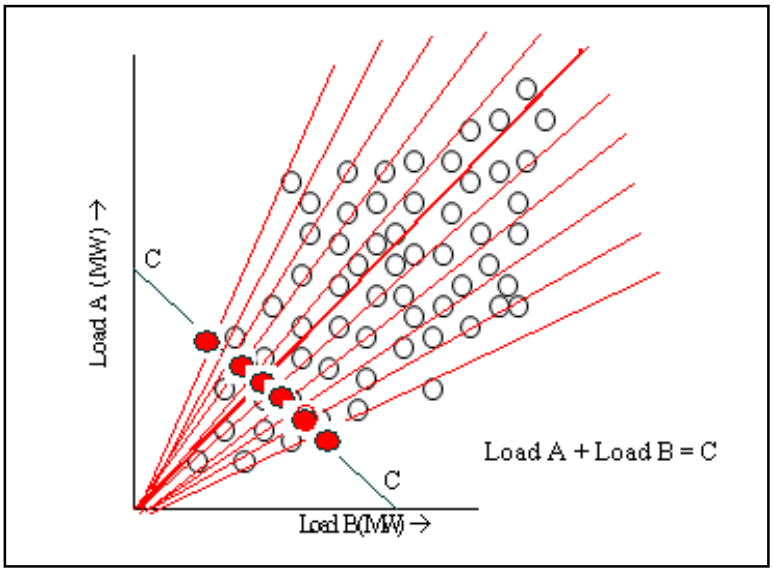


Fig. 3.9 Homothetic stress direction sampling in the load state space

CPF is performed on these basecases along their intrinsic stress directions as shown in the left hand side of Fig. 3.10. This computes the maximum loadability along every stress direction, which is consequently translated into boundary limits, $\{P_{Lmin}, P_{Lmax}\}$ of total system load state space. This limit in the hyperspace is subject to variation due to the influence of discrete variables, i.e., SVC and generator unavailability states. The effect of these two variables is estimated using margin sensitivities with respect to real and reactive power injections along every stress direction, and is given by the equation (3.12),

$$\Delta P_L^{svc} = Q^*_{svc} \,.\, dVSMdQ_{svc} \qquad (3.12)$$

where ΔP_L^{svc} is the change in boundary limit in a particular stress direction due to the influence of SVC unavailability, Q^*_{svc} is the amount of SVC reactive power output at the collapse point along that particular stress direction, and $dVSMdQ_{svc}$ is the linear sensitivity of

voltage stability margin with respect to reactive power injection at the SVC node computed as a by-product of CPF study in that particular stress direction.

Finally, the boundary limits in terms of total system load (MW) identified along every direction can be translated as a boundary region in the total Brittany load state space (univariate distribution as shown in right hand side of Fig 3.10.

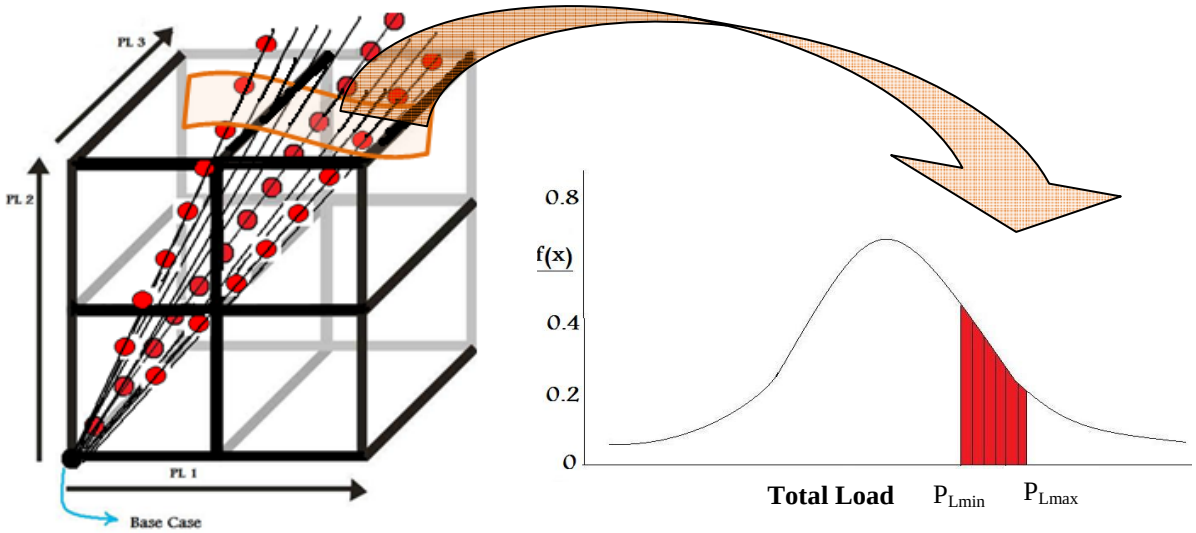


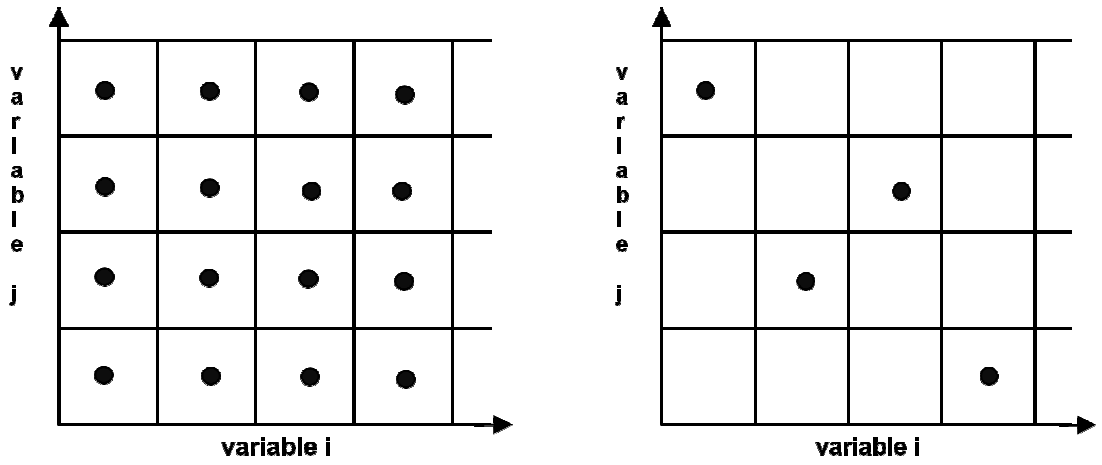
Fig. 3.10 Latin hypercube sampling of stress direction in 3-D and boundary identification

The key in realizing the computational benefit that CPF and linear sensitivity offer lies in the way the homothetic stress directions from the historical data are sampled.

3.3.1.3 Latin Hypercube Sampling of Stress Directions

Latin Hypercube Sampling (LHS) is very prevalently used in Monte Carlo based reliability studies in many fields. LHS of multivariate distribution is performed by dividing every variable forming the multivariate distribution into *k* equiprobable intervals, and sampling once from each interval of the variable. Then these samples are paired randomly to form *k* random vectors from the multivariate distribution. Figure 3.11 depicts the stratified sampling in both forms, traditional and LHS, where the difference is in the pairing process.

In the traditional stratified sampling, samples from every interval of variable *i* is paired with every other samples from all intervals of variable *j*; whereas in the LHS, one sample from an interval of variable *i* is paired only once with any one of the sample from an interval of variable *j*. The pairing in LHS can also be done in such a way as to account for the mutual correlation of the variables by preserving their rank correlation [81], and hence capturing the dependence structure of the multivariate distribution.



(a) (b)

Fig. 3.11 Stratified sampling - (a) traditional, (b) LHS

Similarly, LHS of homothetic stress directions is performed by dividing every stress factor variable obtained from historical data into *k* equidistant intervals (i.e., equal width; a modification to traditional LHS that partitions into equiprobable intervals), sampling once from each interval of the variable, and pairing them preserving their rank correlation, to form *k* homothetic stress directions.

The range of every stress factor variable and their mutual correlation are obtained from the historical data. Figure 3.12 shows a typical stress factor matrix D obtained using

historical data, where each row holds the stress factors of individual loads for a particular historical operating condition. The matrix D is in the form of a multivariate distribution comprised of various vectors of individual load stress factors, which also provides mutual correlation. So LHS is employed to sample random vectors of correlated stress factors that provide us the required stress directions.

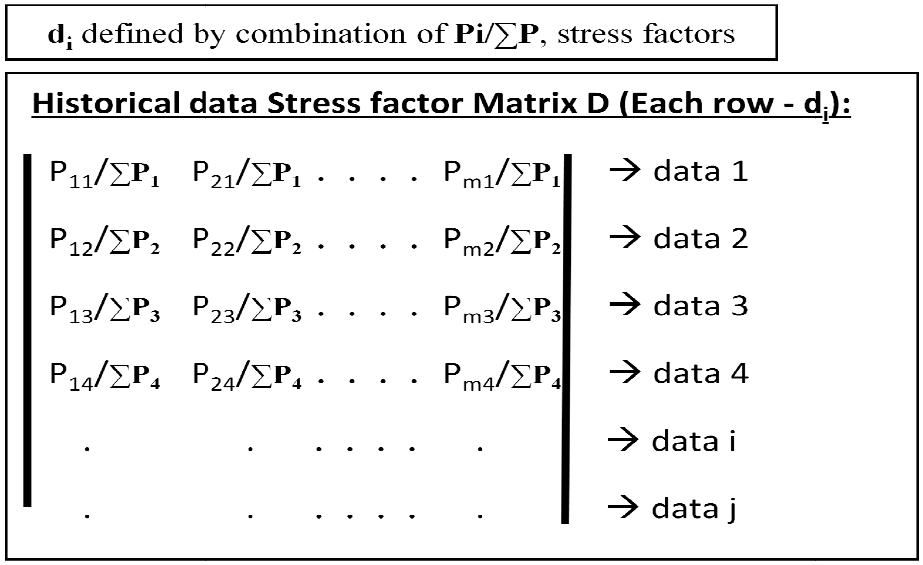
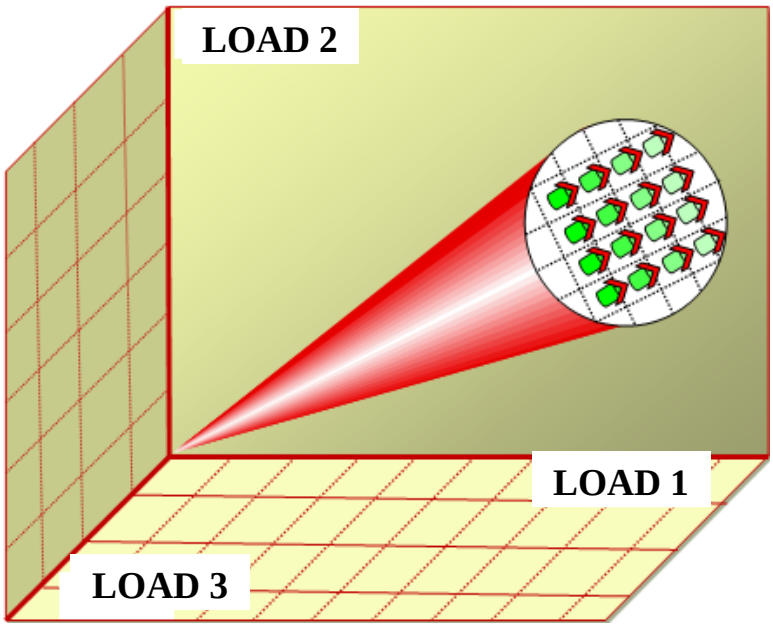
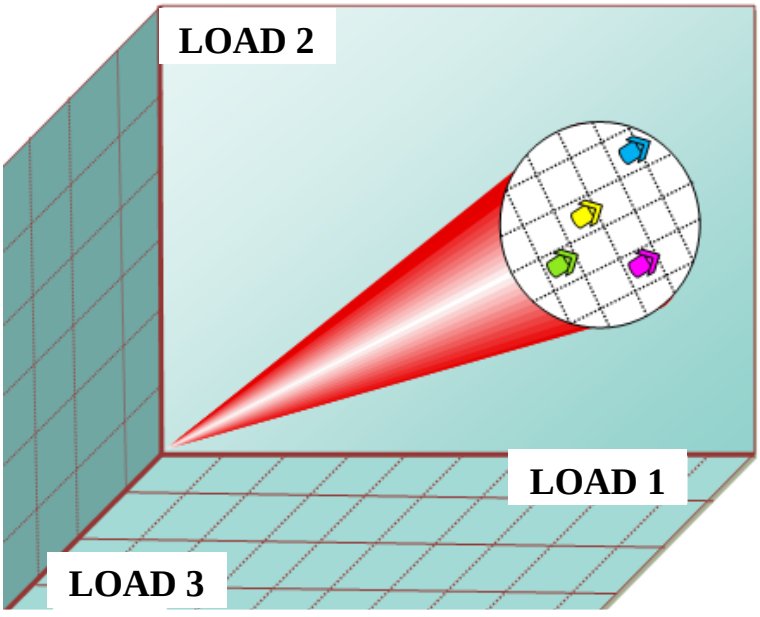


Fig. 3.12 Stress direction defined in terms of stress factors

Figure 3.13 shows (a) traditional stratified sampling and (b) LHS of homothetic stress directions in 3-dimensional state space. In the case of LHS, for *k* intervals per dimension, irrespective of state space size the uniform stratification of stress direction is achieved with *k* samples; compared to stratified sampling that produces k^{n-1} samples for *k* intervals per dimension, in a state space of dimension *n*. The ideal number of *k* is found in an incremental fashion until there is no improvement in the boundary limits. Hence computation to find the boundary region can be decreased drastically by using the proposed method based on LHS of stress directions and linear sensitivities.



a) Traditional stratified sampling



b) Latin hypercube sampling

Fig. 3.13 Sampling homothetic stress directions for boundary identification

3.3.2 Stage II – Sampling

As explained in chapter 2, the property of importance sampling to bias the sampling using an importance function *g(x)* towards the area of interest *h(x)* is used in our method to generate influential operating conditions from load state space *X* with density *f(x)*. So given *S*, the identified boundary region, the importance sampling distribution *g(x)* in general can be constructed as shown in equation (2.7). In the multivariate case, sampling techniques such as copulas or LHS or sequential conditional marginal sampling (SCMS) [71, 82] is used to generate correlated multivariate random vectors from non-parametric distributions $f_1(x)$ and $f_2(x)$. The SCMS method is time consuming and requires a lot of memory usage for storing the entire historical data, while LHS and copulas are relatively faster and consume less memory since they work only with non-parametric marginal distributions and correlation data. We use copulas for their simpler and elegant approach in handling any non-parametric marginal distributions and inter-dependencies. Setting p=0.75, 75% of the points is expected from *N*-dimensional boundary region *S*, as the probability distribution is altered to produce more samples from *S*. Figure 3.14 depicts the probability reorientation by importance sampling process towards the boundary region in a 2-dimensional state space. Again, *p* serves as a sliding parameter that controls the extent of biasing between a completely operational study with *p=1* to investment planning study with *p=0*, as observed in chapter 2.

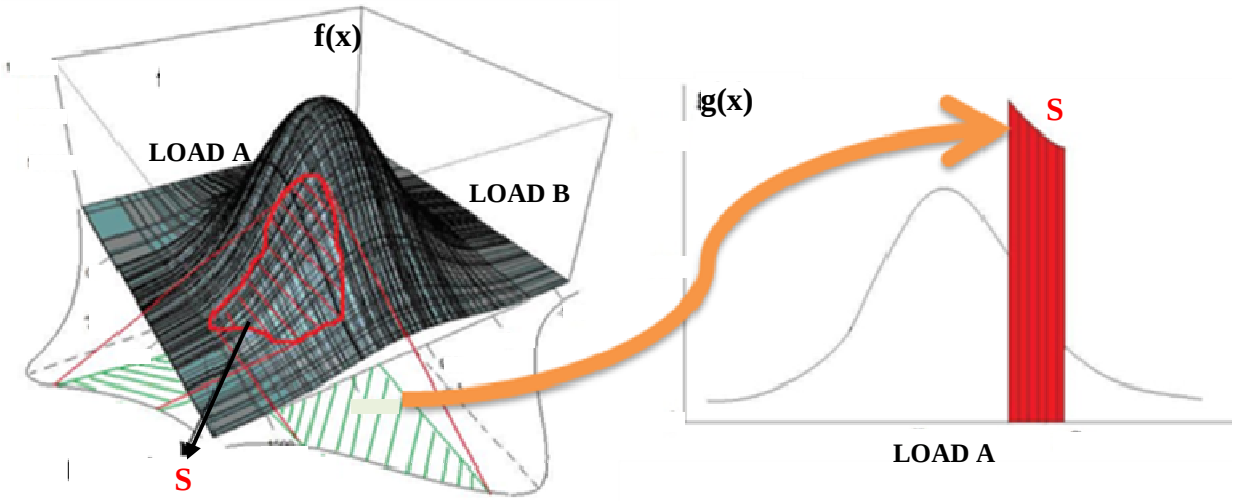


Fig. 3.14 Importance sampling scales up boundary region probability

3.4 Numerical Results

3.4.1 Study Description

The proposed efficient multivariate load data processing approach will be illustrated in a study similar to chapter 2, i.e., a decision tree based security assessment study for deriving operating rules against voltage stability issues on SEO region (*Système Eléctrique Ouest*, West France, Brittany). The following study specifications remain the same as the previous study in chapter 2:

1. The basecase of SEO network considered corresponds to 2006/2007 winter with 13500 MW baseload.
2. The most constraining contingency is the Cordemais busbar fault in the Brittany area that leads to trip nearby group of generation units.
3. Random sampling to generate various basecases is performed on the same set of parameters, i.e., the SEO load, SVC unavailability and generator group unavailability in Brittany area.

4. The sampling laws for the 5 generation units and 2 SVCs remain the same.
5. The simulation parameters, contingency event time, and criteria for labeling scenarios based on post-contingency performance etc., all remain the same.

The major contribution of this study is the consideration of non-parametric nature of multivariate distribution of the system load, with its mutual correlation or inter-load dependency structure preserved, in the efficient Monte Carlo sampling stage.

3.4.2 Data Preparation

As presented in chapter 2, the historical load data during the daytime of winter period (December to February months) between 8hr to 22hr will be used for this study. The multivariate load distribution is comprised of 640 load buses, out of which the data for about 20 load buses were missing completely. While there are maximum likelihood estimation methods such as EM (Expectation Maximization) to iteratively estimate missing or incomplete data, we have used system specific information, i.e., the missing load's proportion to other available loads in the basecase, to estimate the missing load data in the historical records. The following steps explain the method:

Step 1: The ratios of unknown loads (N_{un}) to all other known loads ($N - N_{un}$) in the basecase are calculated; This is refined by including only those known loads that have physical relationship with the unknown loads, such as common control area, region, or any other information that can be obtained from the system experts

Step 2: For a particular historical record, the unknown value of a particular load is estimated with respect to every known load values according to the basecase ratio obtained in step-1. Then the average of all the estimates is considered as the estimation of the unknown load value for that particular historical record. The same is repeated for

every other unknown load values in that historical record.

Step 3: Step 2 is performed for all the missing load values of every historical records.

The reactive power values of the loads are estimated by maintaining the power factor value constant (i.e., basecase power factor). Once the entire historical data consisting of 640 loads is available, the two-stage efficient sampling process can be performed to generate influential operating conditions from the multivariate distribution.

3.4.3 Efficient Sampling of Load Parameter

The proposed efficient sampling method is used to generate samples from the multivariate load distribution obtained from projected historical data.

3.4.3.1 Stage-I: Fast Boundary Region Identification

Performance Measure and Linear sensitivities: The boundary region identification process requires sampling homothetic stress directions using LHS method. The continuation power flow is performed along various stress directions to compute the voltage stability margin, and the computed linear sensitivities are used to estimate the stability margin under the influence of discrete parameter variation. It should be noted that, though actual criteria for declaring a scenario as post-contingency acceptable or unacceptable in the dynamic simulation was based on bus voltage lower limit and simulation convergence status, in the stage of boundary identification stage we propose to use voltage stability margin (which is usually considered as static performance index). Figure 3.15 shows the result of a simulation study performed to validate the above study specification of using voltage stability margin criteria to find the boundary region with respect to voltage collapse, while in the actual dynamic simulation the voltage collapse criteria are different. So two simulation studies were

performed on several operating conditions sampled along the most likely stress direction used in chapter 2:

1. Dynamic simulation using the ASTRE software
2. Voltage stability margin computation using ASTRE

The left hand side of the Fig. 3.15 shows the relationship between the two performance indices, i.e., whenever the simulation doesn't converge before the final time of 1500s, the voltage stability margin computed is less than 0; and whenever the simulation does converge at the final time of 1500s, the voltage stability margin computed is greater than 0.

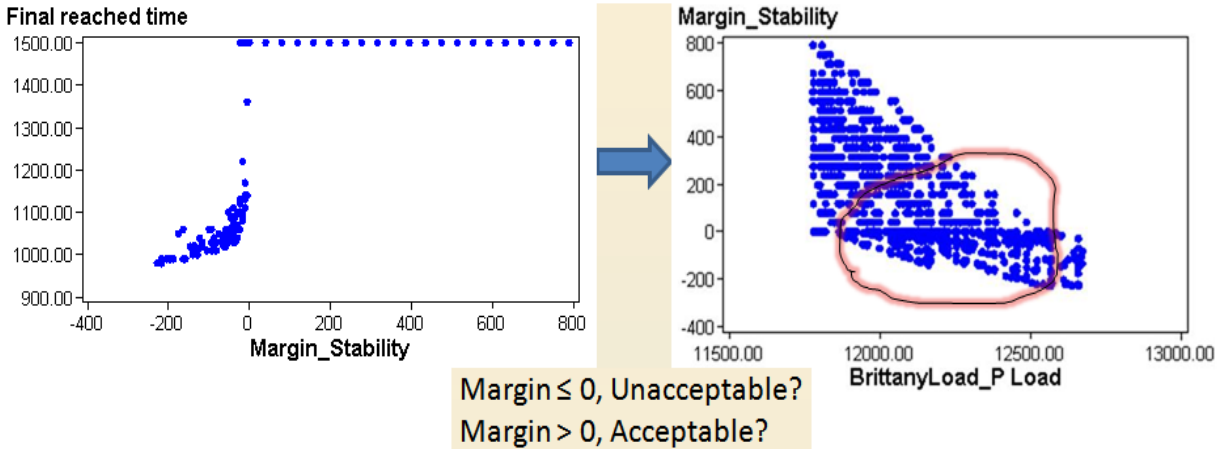


Fig. 3.15 Voltage stability margin as performance index for fast boundary identification

So the following criterion is used to identify the boundary region in the total Brittany load state space, i.e.,

IF VSM≤0, THEN voltage collapse → Unacceptable post-contingency performance

IF VSM>0, THEN NO voltage collapse → Acceptable post-contingency performance

The right hand side of Fig. 3.15 shows the boundary region identified using VSM to be between the same total load limits as was identified in chapter 2 using dynamic simulation

convergence criteria, i.e., 11860 MW and 12600 MW. Hence this corroborates our choice of using VSM and its linear sensitivities to identify the boundary region in the multivariate load state space.

A dynamic simulation study in ASTRE software is performed to identify the voltage stability margin along a stress direction. This computes the collapse point with respect to load increase quickly, as it is a post-contingency process as shown by Fig. 3.16. Unlike the pre-contingency process (left hand side of Fig. 3.16) of performing contingency analysis at every step of system load increase in a particular stress direction and then identify the stability margin at collapse point, post-contingency process of applying contingency and increasing the load until the simulation diverges due to voltage collapse gives the stability margin faster.

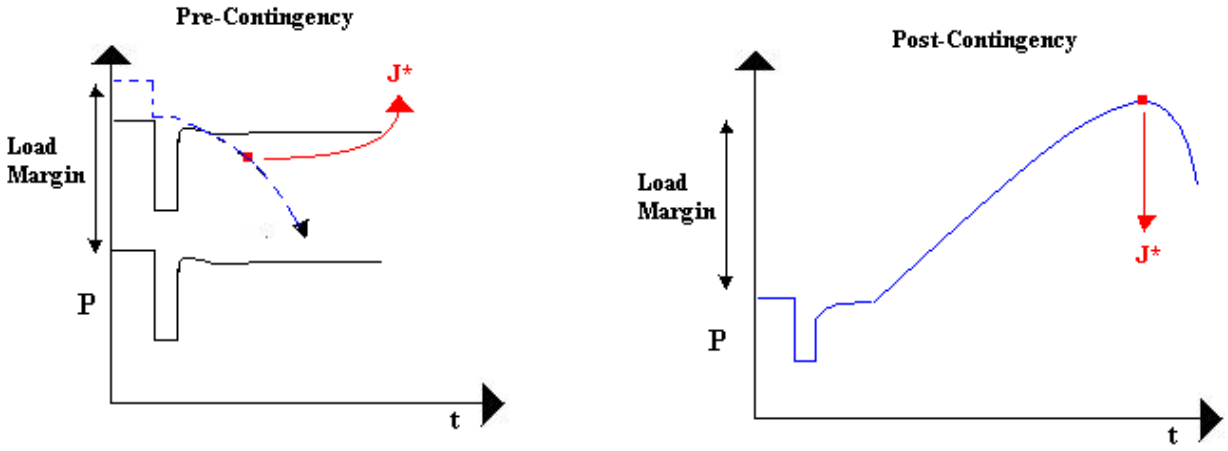


Fig. 3.16 ASTRE simulation options for computing voltage stability margin

Then the power flow jacobian J^* at the collapse point is used to compute the linear sensitivities of VSM with respect to real and reactive power injections, by computing the sensitivity of lowest-voltage bus at the instance of collapse with respect to power injections

at all other nodes [83]. Figures 3.17 and 3.18 show the 400 KV and 225KV voltage results respectively from an ASTRE margin identification simulation done along a particular stress direction on a particular operating condition. The Cordemais bus bar fault was applied at 900s of simulation, and after post-contingency simulation reaches 1500s, the total system load is ramped up at a certain %MW/s along a particular homothetic stress direction considered (i.e., the intrinsic stress direction of the base operating condition under consideration) until the simulation diverges.

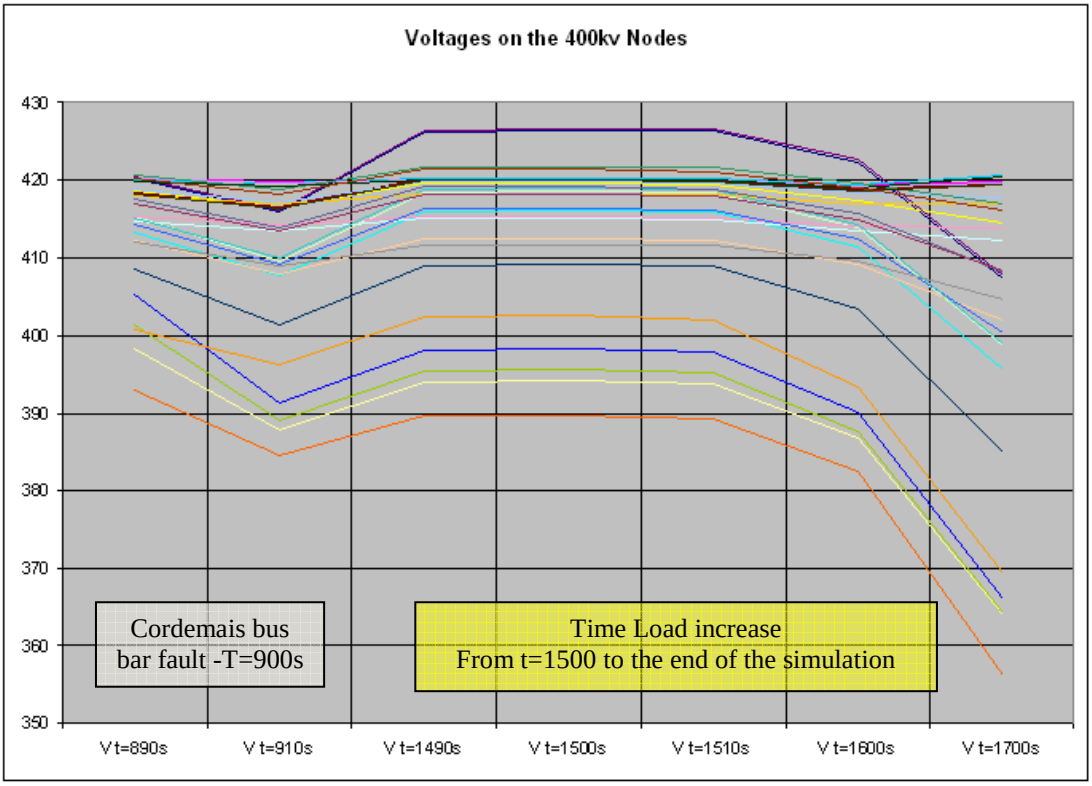


Fig. 3.17 Voltage plots for every 400KV buses

It is noted that the ASTRE simulation diverges at t=1750 s when voltage collapse occurs. The linear sensitivities are computed within ASTRE at this juncture. Likewise, for every sampled stress direction the process of computing voltage stability margin and linear

sensitivities is be repeated in ASTRE. The margin search and sensitivity computation in ASTRE is not as same as the conventional CPF study explained in section 3.3.1, which uses parameterization of system state equations and performs predictor and corrector functionalities iteratively. The boundary identification can also be performed using any other software that finds the bifurcation point and linear sensitivities.

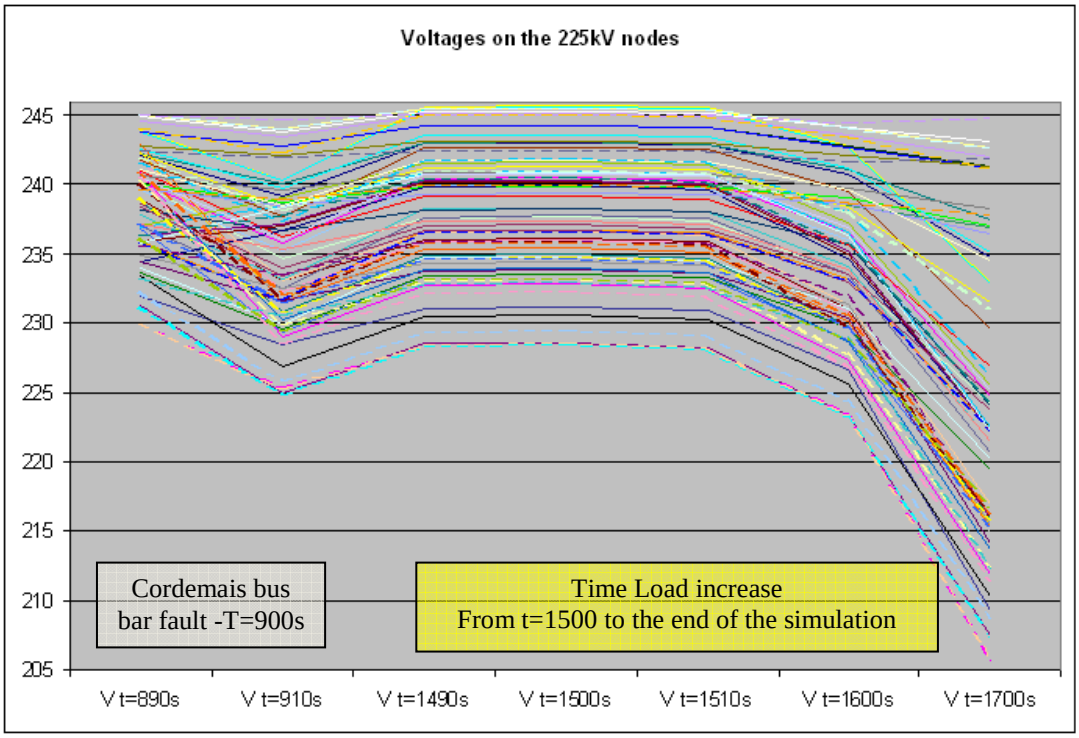


Fig. 3.18 Voltage plots for every 225KV buses

Boundary Identification: There are 24 combinations of discrete parameters (SVC and generator unavailability) as shown in Table 3.1. For the first combination in Table 3.1, with no component unavailability, initial basecases are formed based on the sampled *k* homothetic stress directions. Then CPF is performed to characterize the load state space with respect to post-contingency performance measure and the boundary limits of total SEO load, $\{P_L^{SEO}{}_{min}$,

$P_L{}^{SEO}{}_{max}$} are found, which is {11627, 12700} MW as shown in Table 3.1. Table 3.2 shows the process of estimating *k* for LHS in an incremental fashion. Beyond *k=15*, the boundary region is identified fairly consistently.

Table 3.1 Boundary identification under discrete combinations

S.No	SVC Cases	Generator Cases	$P_L{}^{SEO}{}_{min}$	$P_L{}^{SEO}{}_{max}$
1	None	None	11627	12700
2	None	Blayais	11507	12580
3	None	Chinon	11474	12547
4	None	Civaux	11515	12529
5	None	Flamanville	11476	12506
6	None	St-Laurent	11490	12562
7	Plaine-Haute	None	11618	12691
8	Plaine-Haute	Blayais	11498	12571
9	Plaine-Haute	Chinon	11465	12538
10	Plaine-Haute	Civaux	11506	12520
11	Plaine-Haute	Flamanville	11467	12497
12	Plaine-Haute	St-Laurent	11481	12553
13	Poteau-Rouge	None	11608	12681
14	Poteau-Rouge	Blayais	11488	12561
15	Poteau-Rouge	Chinon	11455	12528
16	Poteau-Rouge	Civaux	11496	12510
17	Poteau-Rouge	Flamanville	11457	12487
18	Poteau-Rouge	St-Laurent	11471	12543
19	Plaine-Haute + Poteau-Rouge	None	11599	12672
20	Plaine-Haute + Poteau-Rouge	Blayais	11479	12552
21	Plaine-Haute + Poteau-Rouge	Chinon	11446	12519
22	Plaine-Haute + Poteau-Rouge	Civaux	11487	12501
23	Plaine-Haute + Poteau-Rouge	Flamanville	11448	12478
24	Plaine-Haute + Poteau-Rouge	St-Laurent	11462	12534
Boundary			11446	12700

The voltage stability margin sensitivities are computed along every *k* stress directions for the basecases with first component combination of Table 3.1. The sensitivities are used to estimate the change in boundary limits due to the influence of component combination change. Table 3.1 also shows the estimated boundary limits for all the remaining

combinations. The final boundary region limits are estimated as 11446 MW (min($P_L^{SEO}{}_{min}$)) and 12700 MW (max($P_L^{SEO}{}_{max}$)).

Table 3.2 Incremental estimation of *k*

k	***P_L^{SEO}min***	***P_L^{SEO}max***	***boundary gap***
5	12500	12700	200
8	11627	12500	873
12	12000	12700	700
15	11627	12700	1073
20	11627	12650	1023
25	11627	12700	1073

Figure 3.19 shows the boundary characterization in terms of total SEO load, obtained from a simulation performed for 24000 random basecases formed by projected historical load data and all combinations of discrete parameters. This result verifies the ability of the proposed method to estimate boundary region approximately at a highly reduced computing requirements (i.e., only about 20 CPF and linear sensitivity computations) in a multivariate parameter state space defined by loads and component unavailability states.

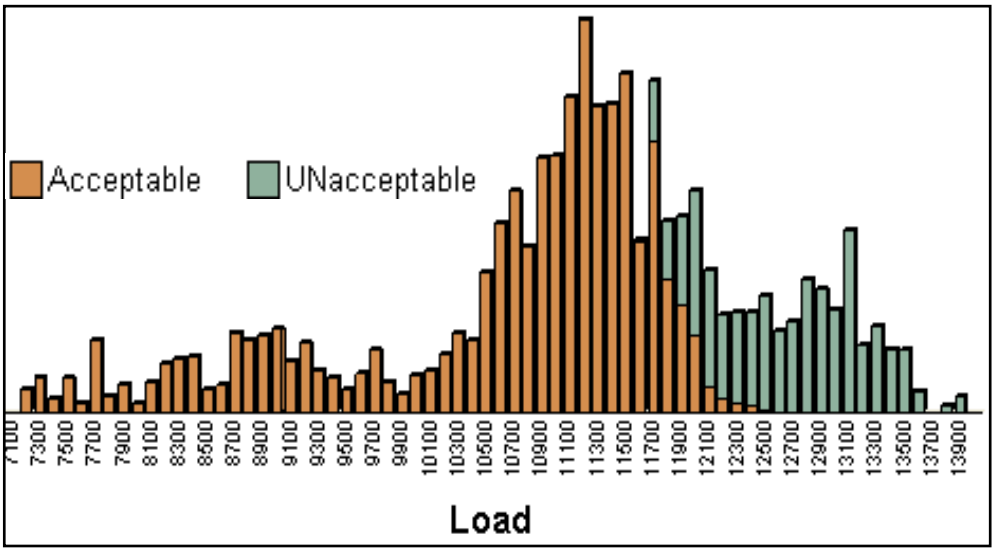


Fig. 3.19 Boundary characterization in total SEO load state space

3.4.3.2 Stage-II: Importance Sampling

Many MCS studies in the past have assumed a multivariate normal distribution of load data [7]. But in our study, we perform importance sampling on actual empirical non-parametric distribution obtained from the projected historical data of loads. Figure 3.20 shows three marginal load distributions among the 640 load vectors that make up the multivariate historical data. It is seen that the multivariate distribution is made up of marginal distributions that are not exactly normal, but by visual inspection some looks close to normal, some uniform, some discrete and so on. So a multivariate Normality assumption may give misleading results.

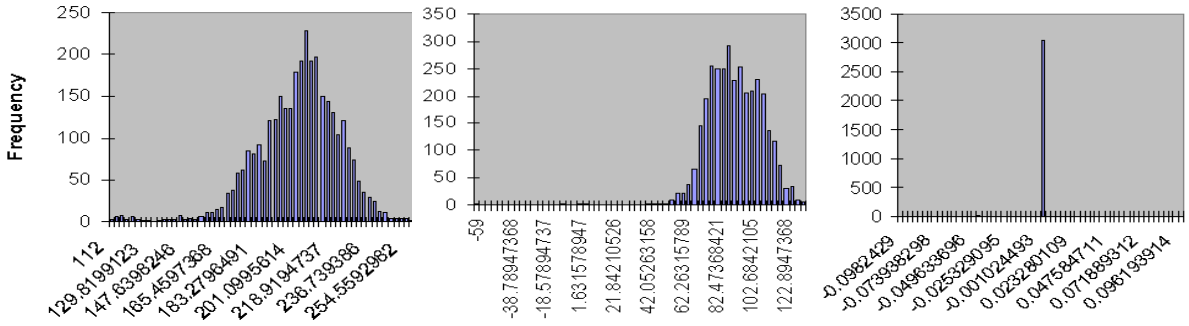


Fig. 3.20 Some sample marginal distributions from historical load data

Furthermore, these marginal distributions are not independent to model them separately as a group of normal, uniform and discrete distributions respectively and sample; but they are mutually correlated, and the sampling method must preserve their inter-dependencies or correlations while sampling. The whole sampling task becomes even more challenging, considering the non-parametric nature of the marginal distributions. Therefore, as mentioned in section 3.3.2, copulas are used that could efficiently work with multiple non-parametric marginal distributions and their mutual correlation (rank correlation) to produce correlated

multivariate random vectors from original multivariate distribution defined by empirical historical data.

After identifying the boundary region limits, the empirical multivariate distribution of boundary region $f_1(x)$ is begotten from historical data by filtering the records within the identified boundary limits. When p = *1* in equation (2.7), we have complete sampling bias towards the boundary region $f_1(x)$. The inter-dependencies between various individual loads are captured in the sampling process by using copulas, and correlated multivariate random vectors from $f_1(x)$ are generated. The generated samples are for real power values only, and the reactive power at the corresponding individual load buses are obtained by maintaining the power factor constant. Figure 3.21 shows the operating conditions sampled in terms of real and reactive load power values from the multivariate boundary region, which is fed as input to ASSESS in the form of a text file.

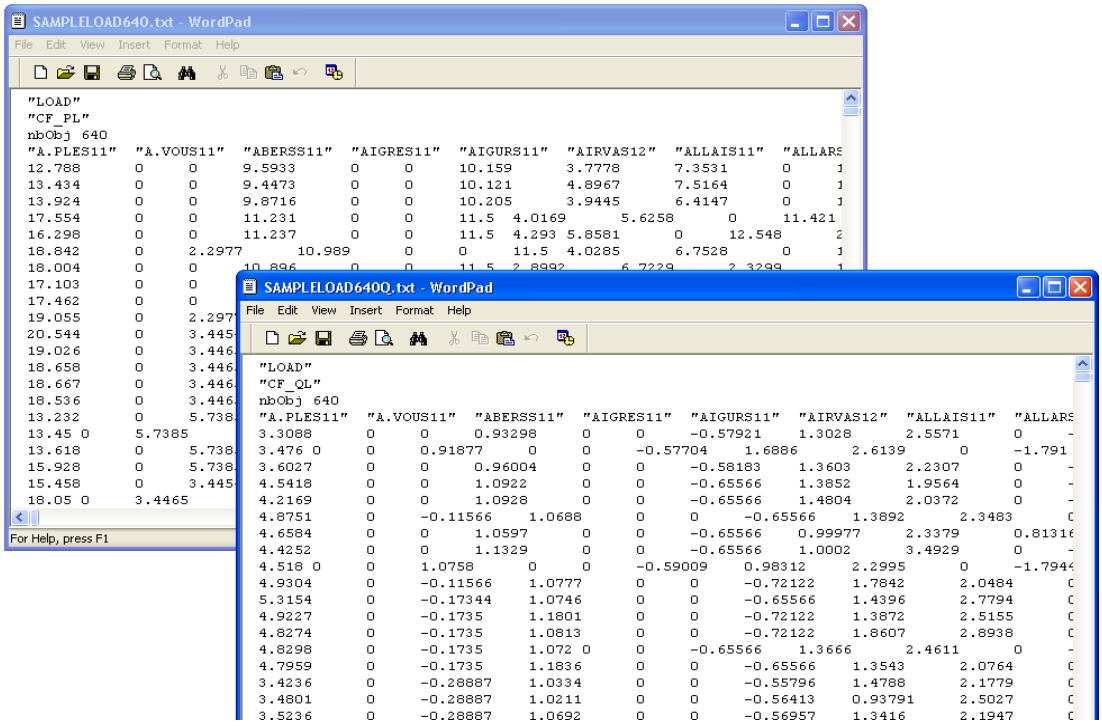


Fig. 3.21 Brittany load samples generated from boundary region importance function g(x)

3.4.4 Results

3.4.4.1 Best Rule Attribute

The training database was generated from the boundary region containing 2852 operating conditions. The test database includes 1976 independent instances, 824 unacceptable and 1152 acceptable cases, covering a wide range of operating conditions unseen by training database. Attribute set “400 KV” contains 46 400KV node voltages, “225 KV” contains 102 225KV node voltages, “P res” contains 10 generator group’s and total SEO real power reserve, “Q flow” contains various attributes such as 12 400KV tie line reactive flows from SEO region to other regions, 4 inter-area 400KV reactive transfers, and net reactive power export; and “Q res” contains 10 generator group’s and total SEO (includes SVCs) reactive reserve. Table 3.3 shows the effectiveness of various attribute sets in terms of classification accuracy and error rates.

Table 3.3 Attribute set selection

Attribute Set	***Accuracy***	***False alarm***	***Risk***	***Tree size***
400 KV + Q res	87.9079	0.193	0.073	15
Q res	87.7159	0.183	0.083	15
225 KV	82.8215	0.243	0.124	15
400 KV + 225 KV	82.7255	0.253	0.12	15
400 KV +225 KV + Q res	82.6296	0.236	0.132	13
All	82.6296	0.236	0.132	13
225 KV + Q res	82.4376	0.231	0.139	13
400 KV	80.8061	0.231	0.166	17
Q flow	75.5278	0.325	0.191	23
P res	73.8004	0.402	0.169	13

Accuracy is defined as the percentage of points correctly classified, false alarm rate is defined as the ratio of total misclassified unacceptable instances among all unacceptable

classifications, and risk rate is defined as the ratio of total misclassified acceptable instances among all acceptable classifications. The attribute set “400KV + Q res” proves to be a good attribute with lowest risk and high classification accuracy. It has to be noted that the accuracy listed in the Table 3.3 are for trees that are pruned by restricting the minimum number of instances per leaf node. On top of this, other dimensionality reduction and attribute selection methods such as principle component analysis, filters and wrappers etc [18], which are very prevalently used in many studies may be employed.

3.4.4.2 Effect of Bias Factor p

Computation, Accuracy and Tree Size: Table 3.4 shows the results when validated using the test database, which confirms that as the sampling of operating conditions is biased towards the boundary region, the entropy of the database increases (a quantitative indicator of information content) and even with lesser database size higher accuracy for decision tree is obtained, also shown in Fig. 3.22. The error rates, namely *false alarms* and *risk*s are both simultaneously reduced to a great degree.

Table 3.4 Performance based on sampling bias

P	***Size***	***Entropy***	***Accuracy***	***False Alarm***	***Risk***
Base	17748	0.7423	92.51	0.063	0.091
0.25	13840	0.7716	93.4211	0.064	0.068
0.50	9932	0.8181	94.9899	0.049	0.051
0.75	6025	0.9038	96.0526	0.038	0.041
1.0	2852	0.9993	97.5202	0.021	0.03

It was also found that as the sampling is biased more towards the boundary region, the size of the decision tree required for good classification also decreases. This is due to the

ability of database to capture high information content (i.e., the variability of performance measure across the security boundary) even with smaller number of instances.

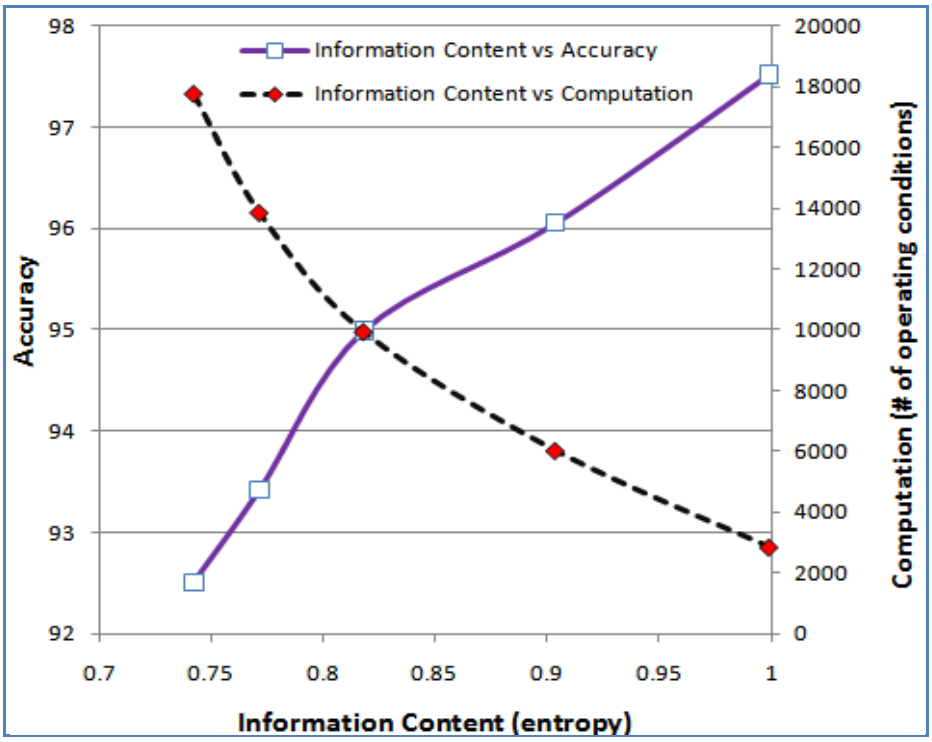


Fig. 3.22 Information content vs. accuracy and computation

Economically beneficial rules: Table 3.5 presents the influence of efficient sampling on the operational rule's ability to provide economic benefit.

Table 3.5 Economic benefit from efficient sampling

Top Node	*p = 0*	*p = 1*
Cordemais voltage	401.64 KV	399.88 KV
Domloup voltage	397.56 KV	394.51 KV
Louisfert voltage	399.1 KV	396.46 KV
Plaine-Haute voltage	392.26 KV	387.21 KV
Chevire unit reactive reserve	131.38 MVar	90.76 MVar
Chinon unit reactive reserve	1127.54 MVar	694.62 Mvar
Cordemais unit reactive reserve	70.97 MVar	16.23 Mvar
Total SEO region reactive reserve	7395.88 MVar	6510.36 Mvar
Plaine-Haute SVC output	11.82 MVar	13.64 MVar
Poteau-Rouge SVC output	16.3 MVar	22.03 MVar

The Table 3.5 shows that for the various possibilities of the decision tree's top node among the most influential attributes, the database generated within boundary region with *p=1* finds rules with attribute thresholds that are always less conservative than from the database generated with *p=0*, i.e., from entire operational state space. Figure 3.23 shows operational rule formed using two attributes, namely reactive reserves at Chevire unit and Chinon unit respectively.

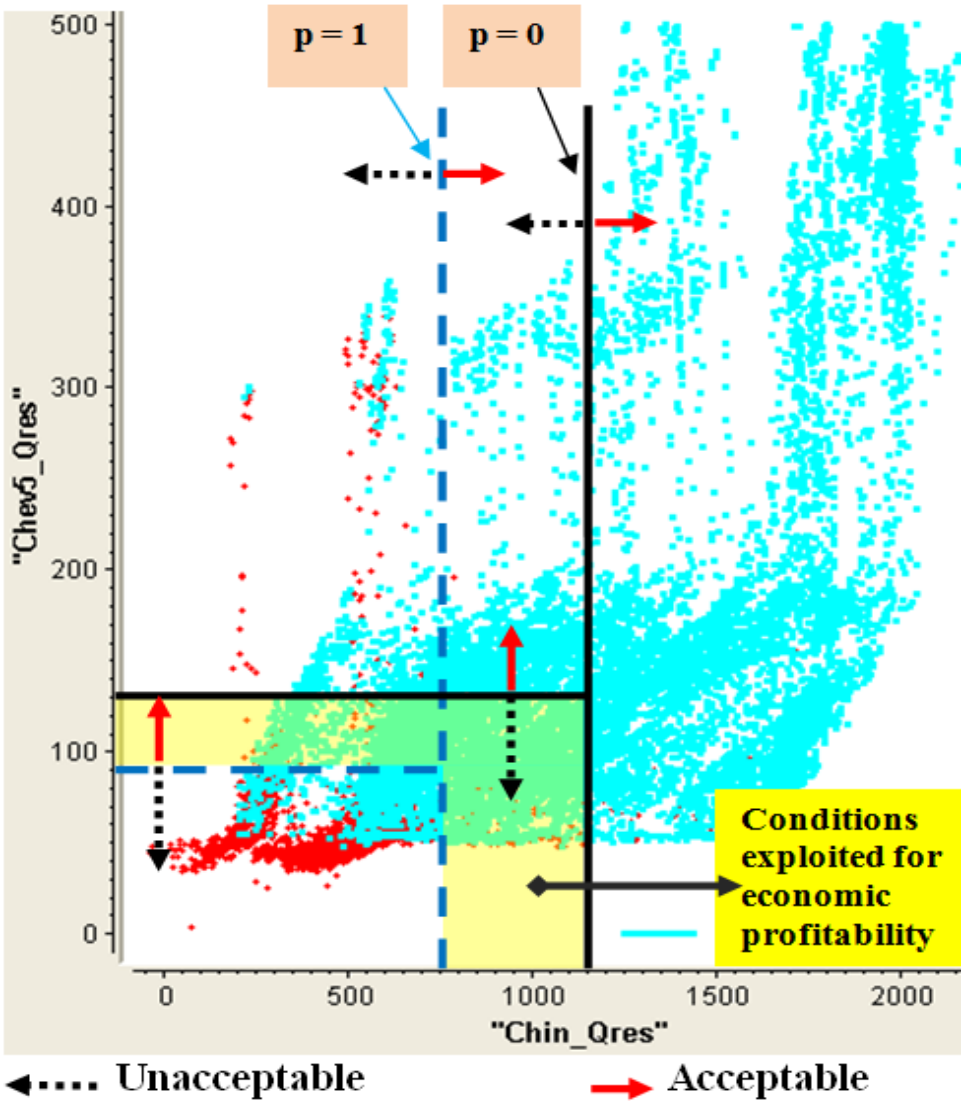


Fig. 3.23 Economical benefit of operational rules from efficient sampling

The operating conditions shown in the Fig. 3.23 are from the entire database. It can be noticed that the rules formed using the database exclusively from the boundary region is providing more operating conditions to be exploited in real time situations, than the rule derived using the database from entire region; because of the increased knowledge and clarity of boundary limits.

3.4.4.3 Sampling Strategies Comparison

Table 3.6 shows the comparison results of different sampling approaches, namely,

1. Uniform sampling of boundary region in the load state space defined along the most likely stress direction.
2. Importance sampling of boundary region in the load state space defined by the most likely stress direction.
3. Importance sampling of boundary region in the multivariate normal (MVN) load distribution (pruned).
4. Importance sampling of boundary region in the correlated non-parametric multivariate load distribution (MVD) (tree pruned).
5. Same as case 4, with tree un-pruned.

Table 3.6 Comparison between different sampling strategies

Sampling Strategy	Size	Accuracy	False Alarm	Risk
1. **Unif (single stress)**	952	56.0729	0.684	0.097
2. **IS (single stress)**	800	63.5628	0.595	0.041
3. **IS (MVN - pruned)**	2879	80.6142	0.142	0.228
4. **IS (MVD - pruned)**	2852	87.0951	0.094	0.178
5. **IS (MVD)**	2852	97.5202	0.021	0.03

It can be seen from Table 3.6 that, importance sampling procedure, even assuming a load state space along a single stress direction, has better performance in terms of high accuracy and low error rates than uniform sampling within boundary. The database produced by importance sampling of correlated-MVD state space definitely shows better performance, of course with a higher computational cost since sampling includes many stress directions. When the trees are pruned for operator's convenience of usage the accuracy decreases, which can be improved using the accuracy-loop as shown in Fig. 2.2. It also performs better than sampling from MVN load space, which is conventional assumption in many studies.

The significance of sampling from correlated-MVD, i.e., capturing the inter-load dependencies, than from MVN is even strongly vindicated by Fig. 3.24 that shows the top 5 critical attribute locations produced by decision trees from respective databases. The contingency event is shown by a red star. The location of 5 critical monitoring attributes as well as their sequence in the tree matters. Compared to MVN, all the 5 top locations found by correlated-MVD sampling strategy are very interesting ones, with the top node being reactive reserve at a big nuclear plant Chinon, the node in the next level of the tree is closer to the contingency location, the next nodes (3 and 4) in the tree deals with the two SVC locations in Brittany and the attribute of node 5 is right at the contingency location.

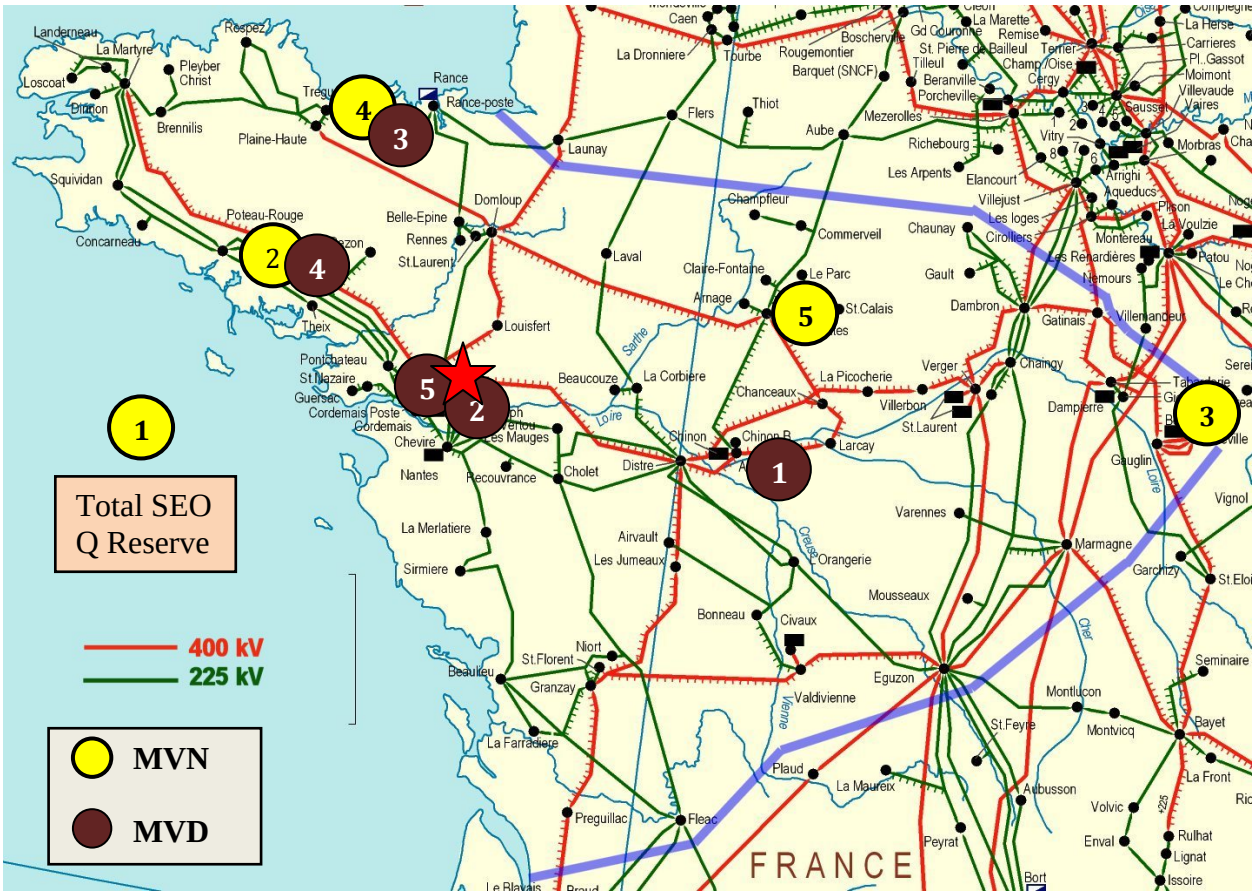


Fig. 3.24 Critical monitoring locations from decision tree: MVD vs. MVN

3.5 CONCLUSIONS

The thrust of the proposed sampling procedure is to re-orient the sampling process to focus more heavily on points for which post-contingency performance is close to the threshold, i.e., boundary region that contains operating conditions influential for rule formation. The chapter emphasizes the significance of sampling from non-parametric correlated-multivariate load distribution obtained from historical data, which ensures selection of attributes from most interesting and relevant locations by decision tree as monitoring locations. A Latin hypercube sampling of homothetic stress direction based linear sensitivity method is developed for quickly characterizing the multivariate load state space

for various combinations of component availabilities, and identify the boundary region with respect to post-contingency performance measure. The developed efficient training database approach was applied for deriving operational rules in a decision tree based voltage stability assessment study on RTE-France's power grid. The results show that the generated training database enhances rules' accuracy at lesser computation compared to other traditional sampling approaches, when validated on an independent test set.

The developed database generation method will also improve the performance of other machine learning classification tools such as SVM, IB*k* etc. The efficient database generation approach can also be applied to other stability problems such as rotor angle stability, out of step etc, where performance measure's trajectory sensitivities will have to be used to reduce computational cost.

This work will have significant benefit to companies owning, operating, or using high voltage transmission systems because it will significantly enhance the speed with which operational planning and investment planning studies are conducted. Companies not familiar with this statistical approach to performing such studies will be interested in the demonstration to gauge its applicability to their own needs.

CHAPTER 4 DECISION TREE BASED SECURITY ASSESSMENT FOR MULTIPLE CONTINGENCIES

4.1 INTRODUCTION

In power system reliability assessment studies the system security limits and adequacy indices depend on the set of contingencies analyzed. Consequently the final solution strategy for short term operational and long term investment planning studies respectively also depend on the set of contingencies considered in the planning study. In chapters 2 and 3, the decision tree based security assessment was performed for the most constraining contingency in Brittany region, which is typically done in many studies. The assumption is that the solution strategy or in our case the operational rules for the most constraining contingency will also perform well on the contingencies that have lower severity. But this is generally not true. In reality, under the highly uncertain nature of power system conditions, the operational rules for the most constraining contingency may not be effective for all other contingencies. Some contingencies, which are generally less severe, may have pronounced ill-effect during certain operating conditions.

For instance, in Fig. 4.1 let us consider an operating condition state space defined by two loads *Pload1* and *Pload2*. Let the two curves (green and orange curves) on the state space indicate the security boundary limits separating the acceptable and unacceptable operating conditions with respect to post-contingency system performance for contingencies 1 (C1) and 2 (C2) respectively. So inducing an operational rule for C1, which is more severe than C2, will classify the operating condition *P* as safe under both the contingencies, while it may not be so. Since the proposed efficient database generation approach in chapter 2 is based on

sampling operating conditions from the boundary region defined by post-contingency performance, now the boundary region has to be defined with respect to multiple contingencies. This will ensure sampling the required high information content training data for decision tree rule formation applicable to multiple contingencies. Therefore, it is important to perform thorough contingency analysis of many contingencies, screen the most important ones that may violate reliability criteria and devise effective solution strategies [84].

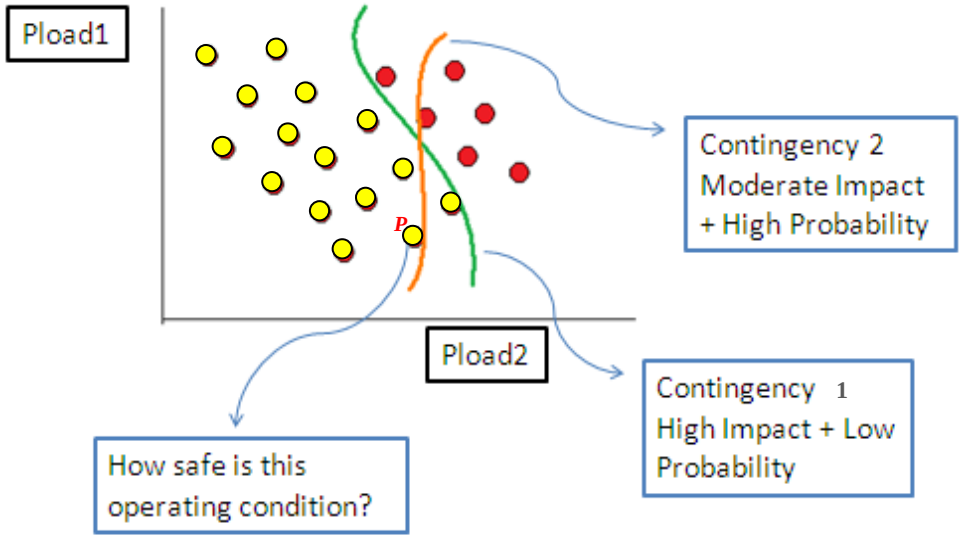


Fig. 4.1 Significance of considering multiple contingencies

So, this chapter focuses on devising efficient methodologies to perform decision tree based security assessment against voltage stability phenomenon for many critical contingencies and obtain operational rules for every contingencies considered. The two main concepts proposed in this chapter for a comprehensive multiple-contingency security assessment are ***risk based contingency ranking*** and ***contingency grouping***.

4.2 MOTIVATION AND PROPOSAL

4.2.1 Risk Based Contingency Ranking

In order to reduce the computational burden of contingency analysis, contingency ranking methods are typically used in power system reliability assessment studies. They help in screening the most critical set of contingencies that are to be thoroughly analyzed. Many deterministic ranking methods have been developed for reliability assessment that considers the severity of contingencies only [58, 59, 85]. While some studies choose the most severe contingency, many screen a credible list of contingencies for planning under a wide range of scenarios. But, under the current highly probabilistic nature of power system, a contingency ranking method which does not consider the probability of each contingency would lead to inconsistent or less effective or even expensive operational solutions strategies. As shown in Fig. 4.1, the C2 even though has a moderate impact or severity on system performance still is highly probable than C1, so it is important give attention to C2 in the planning study. At the same time, there could be some other contingency which has a very severe impact on system performance, but is highly unlikely to occur. In that case such contingencies may be discounted in the overall planning process, or could be considered as some special case independent of overall planning process. Otherwise that one contingency which is very rare, if considered with all other contingencies in the planning process may give forth to very expensive solution strategy in normal operating situations. So we propose to develop *risk based contingency ranking process* that would eventually help in screening top contingencies that may lead to voltage collapse.

The risk of a contingency over a wide variety of operating conditions is defined as,

Contingency *Risk* **=** ***Contingency Occurrence*** *Probability* ***X Contingency*** *Severity* (4.1)

All the risk based contingency ranking process proposed in open literature has the common idea of performing contingency simulations over a wide range of operating conditions, and compute a severity function based upon the post-contingency response database. Then according to the formula shown in equation (4.1), the risk of the contingency is estimated. The same procedure is followed for every other contingency in the selected list, and finally ranked. But the methods developed so far have not considered the actual probabilistic multivariate distribution of the operating conditions, which may also be non-parametric, during the stage of Monte Carlo sampling process. The studies so far have also not considered the huge computational cost incurred in estimating the risk posed by each contingency over many operating conditions. So in this chapter we propose a risk-based contingency ranking method that estimates contingency risk for many contingencies over a wide range of operating conditions sampled from multivariate probability distribution. The proposed method is efficient compared to existing methods in the following, i.e., it has the *ability to get realistic risk indices for multiple contingencies at a very highly reduced computational cost.* The risk indices are realistic because we consider the nature of probability distribution of operating parameters, i.e., if the operating parameter distribution is multivariate normal or it is non-parametric, and efficient methods are developed to address both the situations, which has been missing in all the other works. At the same time, even after accounting for the multivariate nature of operating condition distribution, the risk estimation process is faster as the computation of risk estimation is performed using linear sensitivity information.

4.2.2 Contingency Grouping

Once the critical contingencies have been screened using the risk based contingency ranking scheme, every screened contingency has to be considered for operational planning. Usually, a separate operational rule for every contingency gives the best performance in terms of decision rule's accuracy [60]. So in our study, as shown in Fig. 2.2 we could generate high information content database for every screened contingency, and produce operational rules using decision trees, i.e., in other words, a separate decision tree for every contingency. But this is generally not preferred as it burdens the system operators, who will be dealing with too many rules.

So a global decision tree for many contingencies can be constructed. We could achieve this by sampling operating conditions from the boundary regions of every contingency. But the global tree can never outperform on its ability to classify all the post-contingency situations (i.e., a wider boundary region), when compared to the original separate tree for every contingency. Moreover, there is also the danger of reducing the operating rule's ability to perform well under the most constraining and likely contingency, when we group all the contingencies together. So generally such global trees require usage of decision tree post-processing methods [29] or meta-learning methods such as bagging, boosting, stacking of many learning methods (i.e., divide the boundary region and conquer) [18] etc. to improve its accuracy over the entire domain of boundary region. The problem with these are that they usually overfit the decision tree to the particular operating conditions and contingencies under consideration, and makes the tree very less effective in classifying rare instances. In addition to that, the meta-models do use multiple-trees and voting schemes to classify, and thus it makes the decision process complex for the operators to interpret and apply.

So we propose a contingency grouping method that would strike a balance between producing simple and accurate trees for contingencies, as well as reducing the number of trees for multiple contingencies. The idea of grouping components based on specific performance criteria is already prevalent in power system, as it reduces computational cost for system reliability studies and also provides valuable guidance in decision making. For instance, generators are grouped based on their slow-coherency performance which gives valuable information in controlling islanding to prevent blackout [86]. Generators are grouped based on angle gap criteria for fast contingency screening [87]. Unsupervised learning methods are used to group contingencies based on their effect on bus voltages [88]. Then Neural Networks are used to predict post-contingency bus voltages under many contingencies just by using few representative contingencies, thereby reducing computation. Such grouping concepts are also used for designing defense systems, such as UFLS schemes [89]. So in this chapter, we propose to group contingencies based on the degree of overlapping among post-contingency performances of contingencies over wide range of operating conditions. We introduce a graphical index, termed as *progressive entropy* that captures this degree of overlap visually. The *progressive Entropy curves* are plotted for various contingencies over the distribution of operating conditions along any system variable. The final decision on the potential grouping indicated by *progressive entropy* curves will be based on the particular group's common decision tree's classification performance for all the contingencies in that particular group.

4.3 TECHNICAL APPROACH

4.3.1 Risk Based Contingency Ranking

4.3.1.1 Voltage Collapse Risk of a Contingency

A simple expression for computing risk of a contingency over many probable operating conditions is shown in equation (4.2).

$$\text{Risk}(C_i) = P(C_i) \sum_j P(X_j|C_i) \times \text{Sev}(X_j|C_i)) \tag{4.2}$$

where,

- $P(C_i)$ is the probability of the i^{th} contingency C_i. Assuming that this probability is determined only by the failure rate of the component that causes that contingency, it will be the same for all operating conditions.
- X_j is the j^{th} possible operating condition, and $P(X_j|C_i)$ is the probability of the operating condition given the contingency.
- $Sev(X_{j,}|C_i)$ quantifies the severity of the j^{th} possible operating condition in terms of some stability criteria, when subjected to i^{th} contingency.
- $\Sigma P(X_j|C_i)\ Sev(X_{j,}|C_i)$ quantifies the severity of a contingency computed using its influence over all the sampled operating conditions, X_j.

Typically, Poisson distribution is used to describe the occurrence of an event in a particular time interval. So given an occurrence rate λ of a contingency in a certain time interval, the probability of that contingency happening at least once in that time interval is

$$P(C_i) = \sum_{x=1}^{\infty} P(x) = 1 - P(x=0) = 1 - e^{-\lambda_i} \tag{4.3}$$

where,

- λ is the mean number of events during a given unit of time
- *x* is the number of occurrence

The term $P(X_j|C_i)$ in equation (4.2) can be substituted by the probability of performance index subject to a contingency, $P(PI|C_i)$ [7]. So for a voltage instability problem, probability distributions of performance indices such as maximum loadability $(P(L_m|C_i))$ or voltage stability margin $(P(M|C_i))$ can be used. Voltage stability margin (M) is defined as,

$$M = L_m - System\ base\ load \tag{4.4}$$

So, for voltage instability problem equation (4.2) becomes,

$$\mathrm{Risk}(C_i) = P(C_i) \sum_j P(M_j|C_i) \times \mathrm{Sev}(M_j, C_i)) \tag{4.5}$$

The severity function for an operating condition in equation (4.5) is defined by discrete or continuous function. Typically, if post-contingency margin is non-positive for a particular operating condition, then a voltage collapse will occur. So irrespective of the magnitude of non-positive stability margin, we assume that the consequence of voltage collapse is very severe and generally unacceptable under any condition. So the severity function of an operating condition for voltage collapse is defined as discrete function in equation (4.6).

$$\mathrm{Sev}(M_j|C_i)) = \begin{cases} 1, if\ M_j \le 0 \\ 0, if\ M_j > 0 \end{cases} \tag{4.6}$$

Since the discrete severity function is like an indicator function for collapse, $I(M \le 0)$, the severity function for a particular contingency becomes a probability term, which we refer to as the probability of collapse subject to contingency C_i. It is expressed as,

$$\begin{aligned} \sum_j P(M_j|C_i) \times \mathrm{Sev}(M_j|C_i)) &= \sum_j P(M_j|C_i) \times I(M_j \le 0|C_i)) \\ &= P(M \le 0 \mid C_i), \vee X_j\text{'}s \end{aligned} \tag{4.7}$$

Therefore, for the given discrete severity function, risk in equation (4.5) is rewritten as,

$$\text{Risk}(C_i) = P(C_i) * P(M \leq 0) \quad (4.8)$$

So, to estimate risk of a contingency over a wide variety of operating conditions, we must estimate probability of collapse, i.e., *P(M≤0)* in equation (4.8). This is the bottleneck in contingency risk estimation (CRE) methods. Typically it is done by contingency simulations over various operating conditions produced by Monte Carlo sampling, as in the case of work [90] that samples many operating conditions in the multivariate parameter space defined by border transactions and system loading conditions. But this is very time consuming, especially if it is to be repeated for several contingencies for ranking purposes. Wan et. al [7] in their effort to estimate risk of an operating condition with respect to voltage collapse proposed utilizing linear sensitivity measures to estimate the performance measure (maximum system loadability), which could drastically reduce the computational burden for estimating probability of collapse term. But it assumes the loading conditions to follow a multivariate normal distribution, which is usually not the case in reality. Furthermore, it computes linear sensitivities for only one stress direction, while in reality the multivariate loading distribution will have many stress directions.

In this chapter, we propose a CRE method that considers various stress directions in multivariate load distribution, while utilizing the ability of sensitivity measures to reduce the computational burden. The LHS method presented in chapter 3 is used to sample various homothetic stress directions in the multivariate load parameter state space. We also propose a machine-learning based CRE method in order to account for the influence of non-parametric nature of multivariate load distribution on risk estimates.

4.3.1.2 CRE I: Multivariate Normal Operating Conditions

Let us consider the uncertainty in operating conditions is represented by system loading conditions. The probabilistic nature of system loading conditions is expressed in terms of real power of individual loads, x_i, that forms a 'n' dimensional operational parameter state space X following a multivariate normal distribution as shown by equation (4.9).

$$X = [x_1 \ldots x_n]^T \sim MVN(\mu_x, \sigma_x^2) \tag{4.9}$$

where μ_x is the mean vector $[\overline{x_1}, \overline{x_2}, \overline{x_3} \ldots \overline{x_n}]^T$ representing the mean operating condition, and σ_x^2 is the variance-covariance matrix obtained from historical data. Performing a continuation study on mean operating condition along a particular stress direction in order to assess the voltage stability under a critical contingency, the maximum loadability, μ_{Lm} and the margin sensitivities S_y^p with respect to real and reactive power injections at the critical point can be obtained. Using the margin sensitivities maximum loadability for many other operating conditions defined by individual load variation can be computed as,

$$L_m = \mu_{Lm} + S_y^{p\,T}.(P - \mu_p) \tag{4.10}$$

where P is the parameter vector, which in our case is individual real power and reactive power load at every nodes for various scenarios, given by;

$$P = [X \quad X * r_{qp}]^T \tag{4.11}$$

where r_{qp} is a diagonal matrix with Q/P ratio at every load node, and $X * r_{qp}$ is the reactive power load at every node with a constant power factor. Therefore, P follows a multivariate normal distribution, i.e,

$$P \sim MVN(\mu_p, \sigma_p^2) \tag{4.12}$$

where, μ_p is the mean parameter vector associated with the mean operating condition for which sensitivity information has been found out, and σ_p^2 is the variance-covariance matrix associated with the parameter matrix. In general we can also have other parameters such as generation dispatch, line reactance, shunt susceptance etc.

Since equation (4.10) is a linear transformation of multivariate normal random variable, it can be proved that *Lm* also follows a normal distribution [91].

$$Lm \sim N(\mu_{Lm}, S_y^{p\,T}.\sigma_p^2.S_y^p) \tag{4.13}$$

Voltage stability margin can be defined as,

$$M = Lm - \sum_{i=1}^{n} x_i \tag{4.14}$$

where $\sum_{i=1}^{n} x_i$ is the total system load, X_{Total}. Therefore,

$$M = \mu_{Lm} + S_y^{p\,T}.(P - \mu_p) - \sum_{i=1}^{n} x_i \tag{4.15}$$

Given *X~MVN*, $\sum_{i=1}^{n} x_i$ also follows a normal distribution, i.e., sum of normal marginals (Central Limit Theorem).

$$X_{Total} = \sum_{i=1}^{n} x_i \sim N(\sum_{i=1}^{n} \bar{x}_i, \sigma_{Xtotal}^2) \tag{4.16}$$

where $\sum_{i=1}^{n} \bar{x}_i$ is the sum of mean of each load component (marginal distribution) of X, σ_{Xtotal}^2 is the variance of X_{Total}. Now, voltage stability margin *M*, i.e., performance measure *Y*, also follows a normal distribution.

$$M \sim Y(x) \sim \mathrm{N}\left(\left(\mu_{Lm} - \sum_{i=1}^{n} \bar{x}_i\right),\left(S_y^{p\,T} . \sigma_p^2 . S_y^p + \sigma_{Xtotal}^2\right)\right) \quad (4.17)$$

So the probability distribution of performance measure from probability distribution of operational parameters can be directly obtained, and $P\ (M \leq 0)$ can be computed. Figure 4.2 illustrates the risk calculation procedure for several contingencies, when we have the operating conditions following a MVN distribution.

It is to be noted that the estimation of L_m using sensitivities in (4.10) will be reliable only for the operating conditions along the particular stress direction, d_i under consideration. So as shown in Fig. 4.2 many stress directions are sampled and the probability of collapse is evaluated for every single stress direction for a particular contingency, $P(collapse|C_i,d_i)$. The final probability of collapse for a contingency is computed as,

$$P(collapse|C_i) = \sum P(d_i) * P(collapse|C_i,d_i) \quad (4.18)$$

The degree of variation among all the terms in the above summation is computed and the variance is checked to see if a particular contingency poses a high risk along a particular stress direction, even though the overall risk considering all the sampled stress directions may be low according to equation 4.18. Consequently a separate planning initiative could be implemented for that particular contingency along that particular stress direction.

The probability of sampled stress directions are computed using *k*-Nearest Neighbour (*k*NN) classification method, an instance based machine learning classification method [18, 92, 93]. The following steps are followed:

1. The training data is composed of sampled stress directions, where each stress direction is considered as a separate class (centroid of clusters) and the stress factor components are the attributes

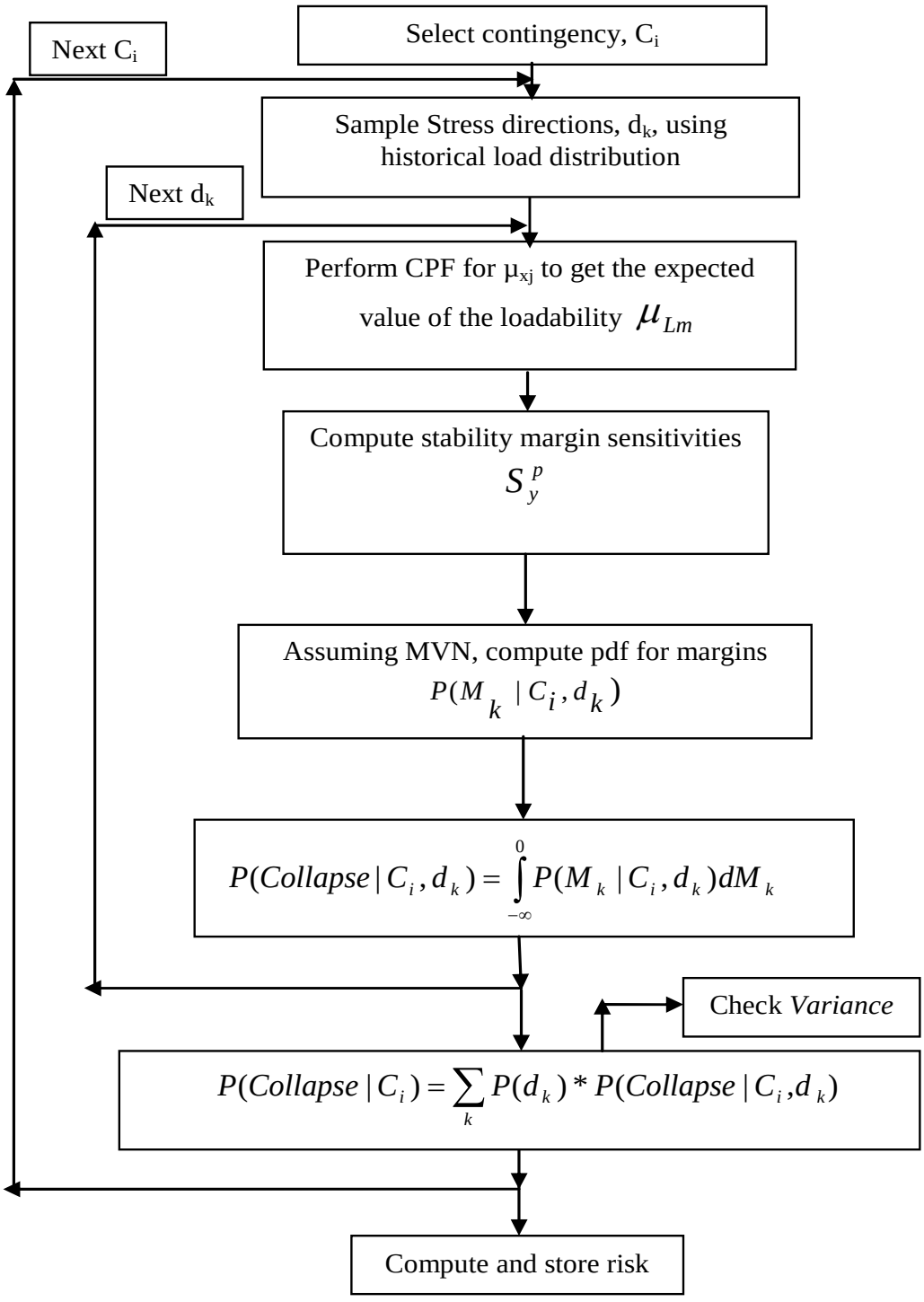


Fig. 4.2 Risk based contingency ranking with MVN assumption

2. The testing data is the stress factor matrix D from historical data

3. The *k*NN classification technique is employed on the training database, and the class predictions for the test database is obtained. In other words, each record in historical stress facor matrix D is mapped onto a particular sampled svector of stress factor using *k*NN
4. Step 3 provides proporion of records in matrix D grouped to each centroid of step 1, and hence the proability of each sampled stress direction is estimated.

Finally, according to equation (4.8), the product of probability of contingency and severity of contingency (probability of collapse) will give the risk of contingency. This is repeated for every selected contingency, their risks are computed and eventually ranked.

4.3.1.3 CRE II: Machine-Learning based Risk Estimation

In section 4.3.1.2, the linear analytical relationship between the operational parameters X and the post-contingency system performance Y (maximum loadability) by virtue of using linear sensitivities as shown in equation (4.10), directly gave forth the probability distribution of post-contingency performance measure for a particular stress direction [7]. This was possible since the operational parameter followed a multivariate normal distribution, which is amenable to linear transformation.

For operational parameter with non-normal or any non-standard distribution, which is usually the case in reality, it is not possible to directly obtain the probability distribution of post contingency performance measure. Therefore Monte Carlo simulation of the operational parameter space X has to be performed to produce many operating conditions, and then the maximum loadability in each case is computed using equation (4.10). This would give the required probability of maximum loadability, which consequently gives the probability of performance measure, i.e., the voltage stability margin M.

It is to be noted here that all the operating conditions sampled from multivariate distribution will not fall in the same stress direction. Hence before using equation (4.10) to estimate the post-contingency performance, we need to compute the linear sensitivity corresponding to the stress direction of particular operating condition under consideration. We can neither afford to compute the linear sensitivities corresponding to the stress directions of all the sampled operating conditions, for it is antidissertation to the very purpose of reducing computation by using linear sensitivities to estimate performance measure. But the fact that operating conditions can be grouped into many clusters based on their proximity of stress directions, can be exploited here to reduce the computation and make effective use of linear sensitivities to estimate voltage stability margin. This is achieved through machine learning techniques.

Figure 4.3 presents the machine learning based risk index estimation method where linear sensitivities computed for few operating conditions are used to estimate the post-contingency performance measure under many other randomly sampled operating conditions. A particular computed sensitivity is associated with a particular new operating conditions based on their intrinsic stress factor vector using *k*NN classification. So the first task is to sample k representative stress directions from the historical data as explained in section 3.3.1.3, for which the maximum loadability and sensitivities are computed beforehand. Then when several operating conditions are sampled, each one is mapped to a particular stress direction among initially sampled k directions using *k*NN classification method. Hence the corresponding sensitivity and loadability values are used in equation (4.10) and the post-contingency performance measure is estimated for that particular operating condition. Likewise, every operating condition is grouped to a particular stress direction, and

accordingly its post-contingency voltage stability margin is estimated using equation (4.14).

The probability of collapse in equation (4.8) is computed using the estimated voltage stability margins for all the 1000 sampled operating conditions, as shown by equation (4.19).

$$\Pr(M \le 0) = Risk\ of\ collapse = \frac{\#(M_i \le 0)}{\#operating\ conditions} \tag{4.19}$$

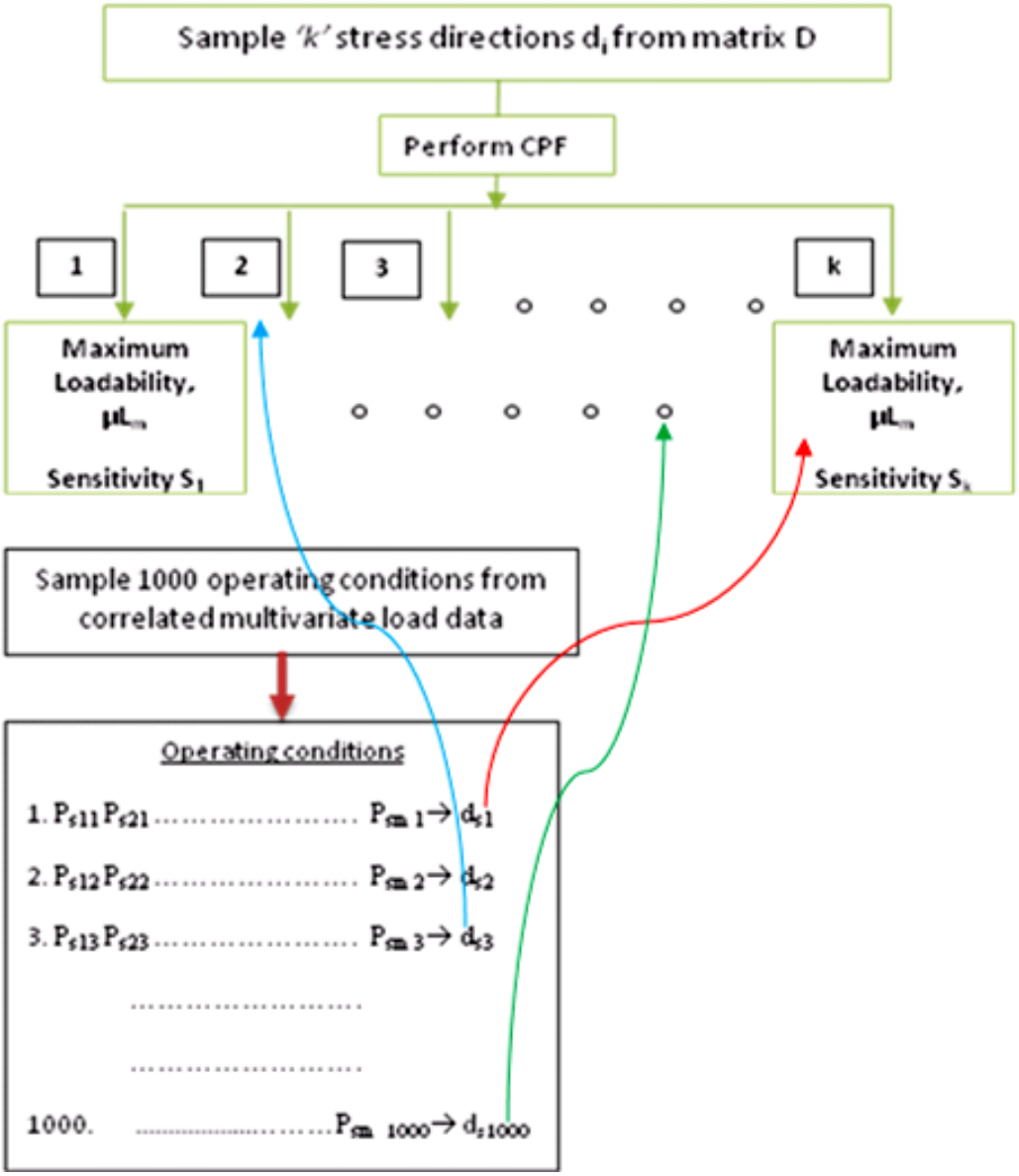


Fig. 4.3 Mapping operating conditions to stress directions using *k*NN classification

Hence the risk of contingency is estimated. The same is done for other contingencies too, and eventually a risk-based contingency ranking is performed.

4.3.2 Contingency Grouping

This section explains the proposed progressive entropy based contingency grouping concept. This is developed to derive a smaller set of rules with good performance for all the screened contingencies. The concept of entropy was discussed in chapter 2, where entropy provides a quantitative measure of information content in a database, i.e., the non-homogeneity level in the class attribute (performance measure) of the database. Here, we introduce a new concept, namely progressive entropy, for visualizing the variability in class attribute along any power system variable, such as system load level, reactive reserve in an area, line flows, generator group reactive reserve etc.

4.3.2.1 Progressive Entropy

Progressive entropy is computed as follows:

Step 1: Sample many operating conditions from the multivariate load distribution

Step 2: Perform simulation and ascertain the post-contingency performance measure

Step 3: Stack the performance measure variability along a system variable distribution. Figure 4.4 shows the boundary progression in the total load variable.

Step 4: Compute the database entropy for every progressive database S_j as shown in equation (4.20), and plot the progressive entropy along any important variable. Figure 4.4 shows the progressive entropy curve for a contingency in the system load variable.

$$\begin{aligned} \text{Pr}\,ogressive\ Entropy &= Entropy(S_j),\ j = 1,2,\text{K}\ N \\ &= \sum_{i=1}^{c_j} - p_i \log_2 p_i \end{aligned} \tag{4.20}$$

where,

- S_j is the progressive database, made up of operating condition x_j's taken one at a time in the direction of going towards unacceptable conditions. So variables such as total Brittany load the unacceptable operating conditions proliferate in ascending direction, and for variables such as reactive reserve the unacceptable operating conditions proliferate in descending direction.
- N is the total number of operating conditions and consequently the total number of progressive databases,
- c_j is the number of classes in the database S_j, and
- p_i is the proportion of S_j classified as class i.

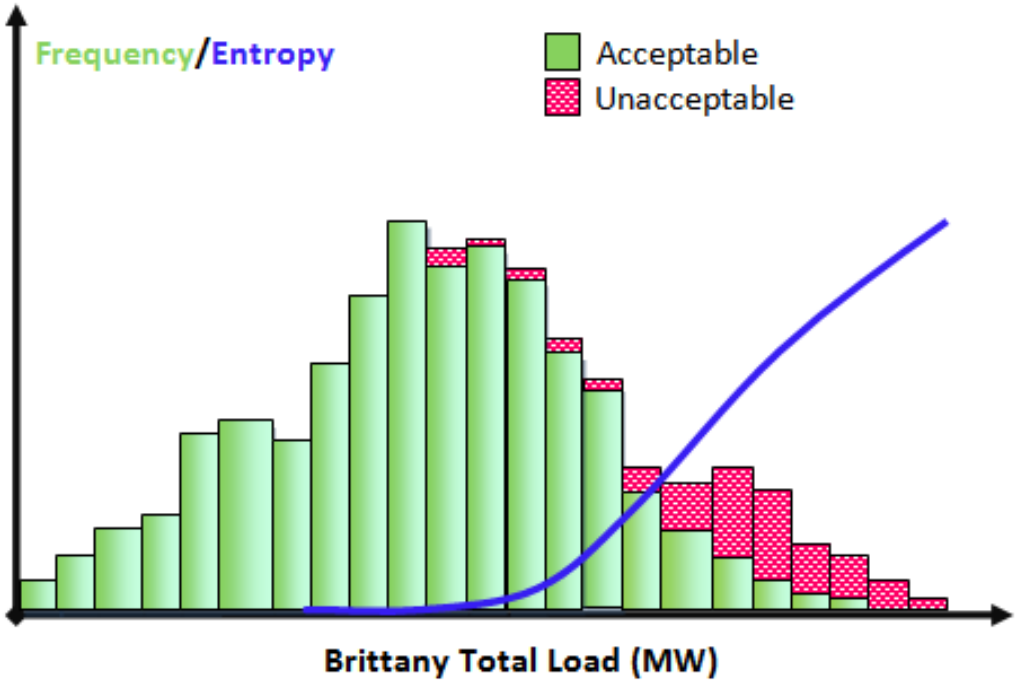


Fig. 4.4 Boundary progression and progressive entropy in total load variable

Again in this case, computational cost can be tremendously saved by using linear sensitivities of performance measure with respect to sampling parameters, i.e., loading conditions, as described in section 3.3.1.1. In this way, we can skip the step-2 mentioned above to compute progressive entropy.

4.3.2.2 Contingency Grouping Recommendations

Figure 4.5 shows the typical progressive entropy curves for 4 different contingencies C_1 (highest risk), C_2, C_3 and C_4; based on which recommendations for contingency grouping will be made.

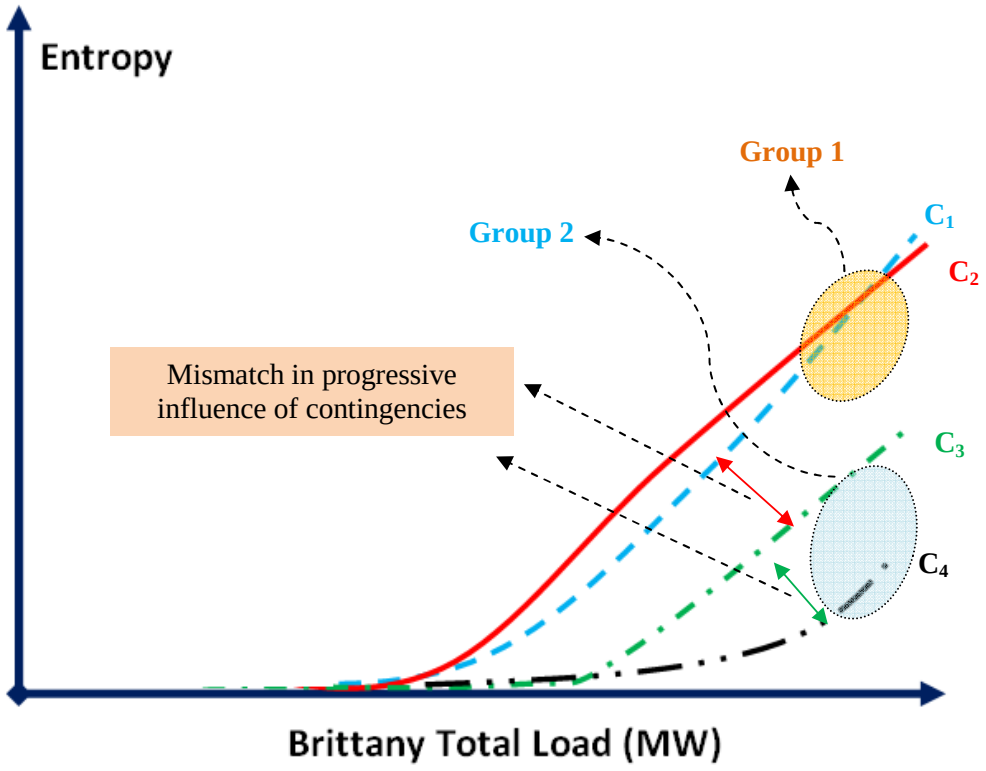


Fig. 4.5 Contingency grouping recommendations based on progressive entropy

The following are some factors that help us in making the decision:

1. The degree of closeness among curves, i.e. whether intertwined or closely enveloping?

2. Visualization of progressive influence of contingencies over operating conditions

For instance, in Fig. 4.5 the progressive entropy curves for C_1 and C_2 along load variable intertwine, indicating they have similar influence on the operating conditions in all the load ranges. So they can be grouped together as Group1 to generate a common operating rule, which is advantageous for the operators. There are two options for generating a training database for a common rule:

1. **High Risk:** The training database is generated by sampling the operating conditions from the boundary region of the contingency that has highest risk among the grouped ones having similar severity. This is to ensure that the rule performs exclusively well for the high risk contingency.
2. **Proportional Risk:** The training database is generated by sampling operating conditions from each contingency's boundary region proportional to its risk index. This is done to bias the training database according to the likelihood of contingencies among the group of contingencies that have similar severity.

Also in Fig. 4.5, Group1 contingencies envelope C_3 and C_4. Similarly C_3 envelopes C_4. But they are not as close as C_1 and C_2, implying that the progressive influence of the contingencies in Group1 over the operating conditions is more severe than the contingencies C_3 and C_4. So if it is not too close, then a common rule may poorly perform. The reduction in performance may manifest in different ways depending upon the choice of training database. For instance, a common rule derived from the training database generated based on the "high risk" criteria will degrade the performance for other contingencies in the following way. The rule will generate a lot of false alarms for less severe contingencies or more risks for more severe contingencies. For instance, the rule for C_1 may produce a lot of false alarms when

applied to C_3 and C_4. If the common rule is generated based on the criteria of “proportional risk”, then there is a great chance of degrading the rule performance for high risk contingency, as the rule has to cater to a wide spread boundary region.

Nevertheless, in the case of less severe contingencies C_3 and C_4, inspite of the above mentioned possible degradations in rule performance, they can still be grouped together as Group2. In this case, the reduction in rule performance generally is very less, since they fall in the lower severity band with smaller boundary regions.

So for each group recommendation, two training databases are generated, i.e., as per high risk and proportional risk criteria. The final common rule is selected based on its performance over all the contingencies in the group. Therefore, the proposed contingency grouping concept promises:

1. Reduction in operating rules. For the hypothetical case considered in Fig. 4.5, rules reduced from five to two for a total of five contingencies.
2. Computation reduction for generating training databases. This is possible since the group recommendations are made prior to the stage of training database generation by using linear sensitivities to obtain progressive entropy curves. will reduce the computational cost involved in generating training database for decision tree training. In the hypothetical case discussed above in Fig. 4.5, four databases are required to derive common rules, instead of five for individual contingencies.
3. Improvement in rule performance by producing a separate common rule for different groups, which is better than overfitting a global common rule for all the contingencies.

4.4 NUMERICAL RESULTS

4.4.1 Risk Based Contingency Ranking

4.4.1.1 Study Description

The proposed risk based contingency ranking approach is applied for a voltage stability study performed on SEO region (*Système Eléctrique Ouest*, West France, Brittany) of French EHV system. Figure 4.6 shows a map of critical contingency locations in French network that are selected in consultation with RTE engineers. These contingencies are usually considered to have severe influence on voltage stability of SEO network during winter. The objective is to rank the considered contingencies in decreasing order of their voltage collapse risk. Eventually the top contingencies are screened, and decision rules derived as per methods proposed in chapter 3.

The details of each contingency in the locations shown in Fig. 4.6 are presented in Table 4.1. The Chinon node, which is not shown in the French network of Fig. 4.6, is near the Avoine node. At Flamanville node, there are two critical units and therefore three different contingencies, i.e., unit 1 outage, unit 2 outage and outage of both the units, are investigated as shown in Table 4.1. Out of seven contingencies considered for the study, three critical ones (at Chinon, Cordemais and Domloup) are within the SEO region, and the rest (at Flamanville and Launay) are outside SEO. Those contingencies outside SEO region, especially Flamanville with two important generators, fall in the western belt of the French network and are considered to impose serious influence on SEO region’s voltage stability performance during heavy transactions.

Fig. 4.6 French EHV network – contingency list

Table 4.1 also provides unavailability rates per year for every contingency. The probability of contingency is computed as per the equation (4.3).

Table 4.1 Contingency probability

Contingency	Unavailability rates/year	Unavailability rates/3 months	Probability
CHINON unit 3	0.1925	0.048125	0.04698
CORDEMAIS bus bar	0.316	0.079	0.07596
DOMLOUP bus bar	0.02235	0.005588	0.00557
FLAMANVILLE unit 1	0.1925	0.048125	0.04698
FLAMANVILLE unit 2	0.1925	0.048125	0.04698
FLAMANVILLE N-2	0.03705	0.002316	0.00220
LAUNAY bus bar	0.02235	0.005588	0.00557

4.4.1.2 Contingency Severity for Single Stress Direction

Table 4.2 presents the results of computing severity function for Cordemais bus bar fault using both the proposed methods (Normal as well as M/C learning), along 10 different stress directions that are sampled using LHS method. The different homothetic stress directions are sampled on the basis of stress factor matrix D obtained from historical data, as was explained in section 3.3.1.3. The probability of the sampled stress directions are computed using the *instance based learning*, *k*NN method explained in section 4.3.1.2, and the results in the Table 4.2 are presented starting from highest probable stress direction to the lowest among the 10 sampled directions.

For the study with the assumption of Normal distribution of loading conditions, the probability of collapse was computed along every single stress direction. In the case of M/C learning method assuming non-Normal distribution, the probability of collapse is estimated by mapping all the sampled operating conditions as shown in Fig. 4.3 to the single stress direction under consideration. It is seen that the estimated contingency severity varies along every stress direction for both the cases.

Table 4.2 Cordemais contingency severity estimation for various stress directions

Stress Direction No.	***Probability of Stress Direction***	***Severity***	
		Normal	*M/C learning*
1	0.24513	0.07509	0.12103
2	0.22667	0.16468	0.17641
3	0.16821	0.09783	0.17436
4	0.14667	0.18423	0.20205
5	0.05231	0.12722	0.26462
6	0.02974	0.12681	0.18154
7	0.01231	0.05548	0.06154
8	0.00513	0.05548	0.10974
9	0.0041	0.19641	0.27282
10	0.00103	0.22757	0.13436

Figures 4.7 and 4.8 show contingency severity results given in Table 4.2 for the decreasing order of stress direction probabilities for Normal and non-parametric assumptions of state space respectively. It is observed that for less likely stress directions, the severity of contingency is very high, as it is true that for rare operating conditions the system is more prone to post-contingency voltage collapse. If we consider the first 6 stress directions, in both Figs. 4.7 and 4.8 it is seen that though stress direction 1 has high probability of occurrence than the stress directions 2, 3, 4, 5 and 6; the severity for later directions are much higher than that of direction 1. So it is important to consider the influence of multiple stress directions over contingency severity estimates. This would ensure proper realistic estimation of risk of contingency over many operating conditions sampled from a multivariate load state space. Otherwise, we will get misleading results.

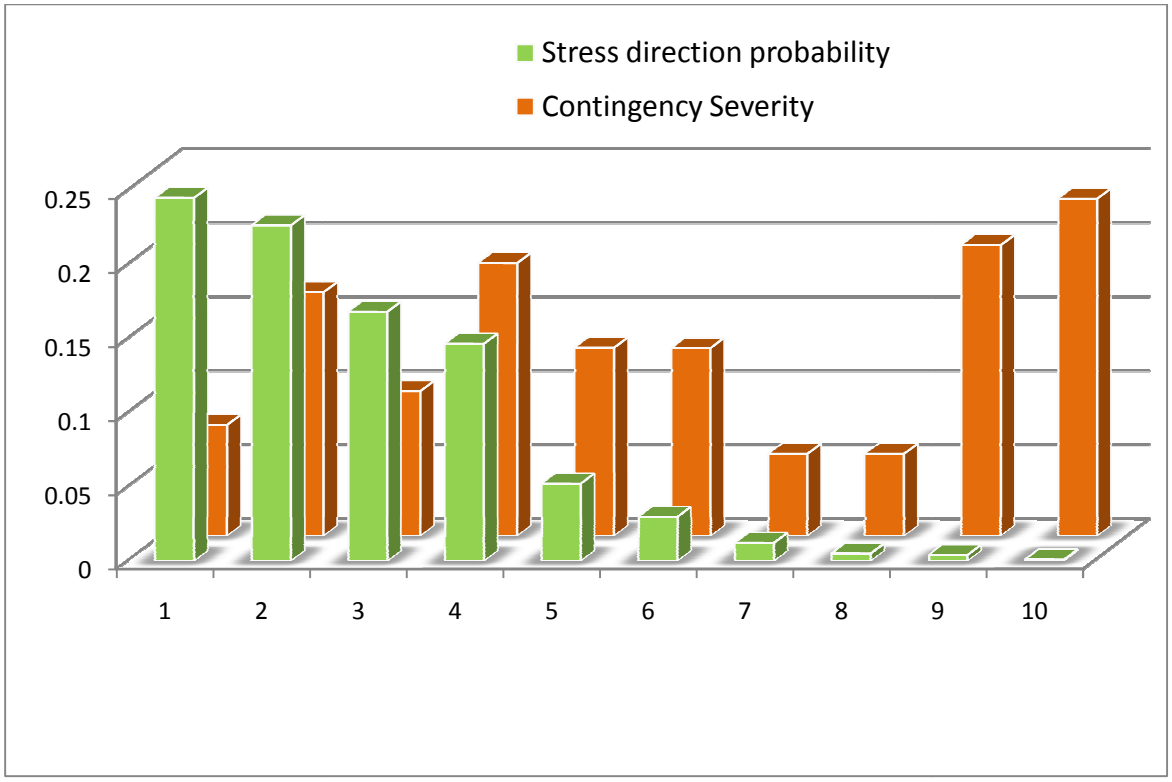


Fig. 4.7 Severity estimation for various single stress directions – MVN assumption

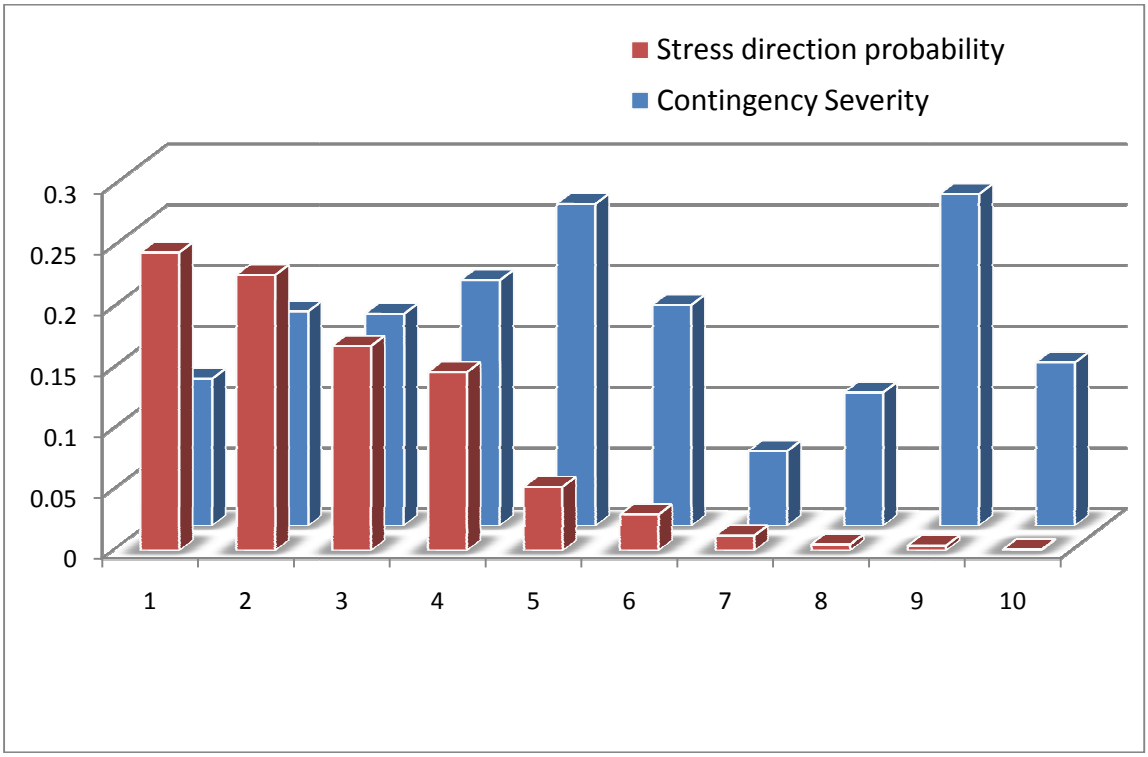


Fig. 4.8 Severity estimation for various single stress directions – M/C learning

4.4.1.3 Contingency Severity for Multiple Stress Directions

Table 4.3 shows the results when multiple stress directions are considered for estimating contingency severity over a multivariate operating parameter state space. The results for three contingencies are shown, for which the severity (probability of collapse) was also computed by performing proper dynamic simulation using ASTRE software. This was done by sampling 975 operating conditions from the non-parametric multivariate load distribution using the copula method explained in chapter 3, which also captures the inter-correlation among various loads. Then the various base cases formed are subject to all the three contingencies systematically using dynamic simulation and the post-contingency performances are analyzed. Using the same post-contingency criteria mentioned in earlier chapter for dynamic simulation, i.e., 400 KV voltages and simulation convergence status, the various base cases are labeled as acceptable or unacceptable; which gives the probability of collapse estimation from simulation. The probability of collapse values estimated by simulation is 0.1702, 0.7446, and 0.1466 for the indicated contingencies at Cordemais, Flamanville and Launay respectively.

Table 4.3 Severity estimate comparisons

S. No	Contingency	Severity			
		SSDS	MVN	M/C	Simulation
1	CORDEMAIS bus bar	0.07509	0.11955	0.16821	0.1702
2	FLAMANVILLE N-2	1.00000	0.80256	0.77128	0.7446
3	LAUNAY bus bar	0.07042	0.09647	0.16000	0.14666

Table 4.3 also shows the contingency severity estimated using various stress directions in three different ways, i.e., SSDS – only considering the most likely single stress direction,

MVN – assuming a multivariate Normal distribution of loading conditions, and M/C – using Machine learning for operating conditions defined by correlated multivariate loads that follows a non-parametric distribution. For MVN and M/C *k=15* different stress directions were sampled in the multivariate state space. It is seen that the estimated results using M/C corroborates with the simulation results. This is due to the fact that the simulation was performed on operating conditions that were sampled from realistic multivariate distribution of load that follows non-parametric distribution with mutual load correlation. Though, MVN study improves on the estimates computed by SSDS closer to the simulation results, nevertheless this study emphasizes that it is essential to take into account the original historical load distribution's characteristics to obtain realistic results. So the proposed M/C based contingency risk estimation method accomplishes this requirement with a very low computational cost.

4.4.1.4 Risk Based Contingency Ranking

Table 4.4 shows the final risk based contingency ranking result for the considered seven contingencies using the proposed M/C method.

Table 4.4 Risk based contingency ranking

Rank	**Contingency**	**Pr (C)**	**Sev (C)**	**Risk (C)**
1	CORDEMAIS bus bar	0.07596	0.1682	0.01277
2	FLAMANVILE unit 2	0.04698	0.2010	0.00944
3	FLAMANVILE unit 1	0.04698	0.1928	0.00905
4	CHINON unit 3	0.04698	0.1077	0.00505
5	FLAMANVILLE N-2	0.00220	0.7713	0.00170
6	LAUNANY bus bar	0.00557	0.16	0.00089
7	DOMLOUP bus bar	0.00557	0.0831	0.00046

Figure 4.9 shows that the contingencies with high severity measure necessarily doesn't pose high risk, and hence a risk based contingency ranking that also accounts for the probability of contingency occurance will be suitable for operational planning study.

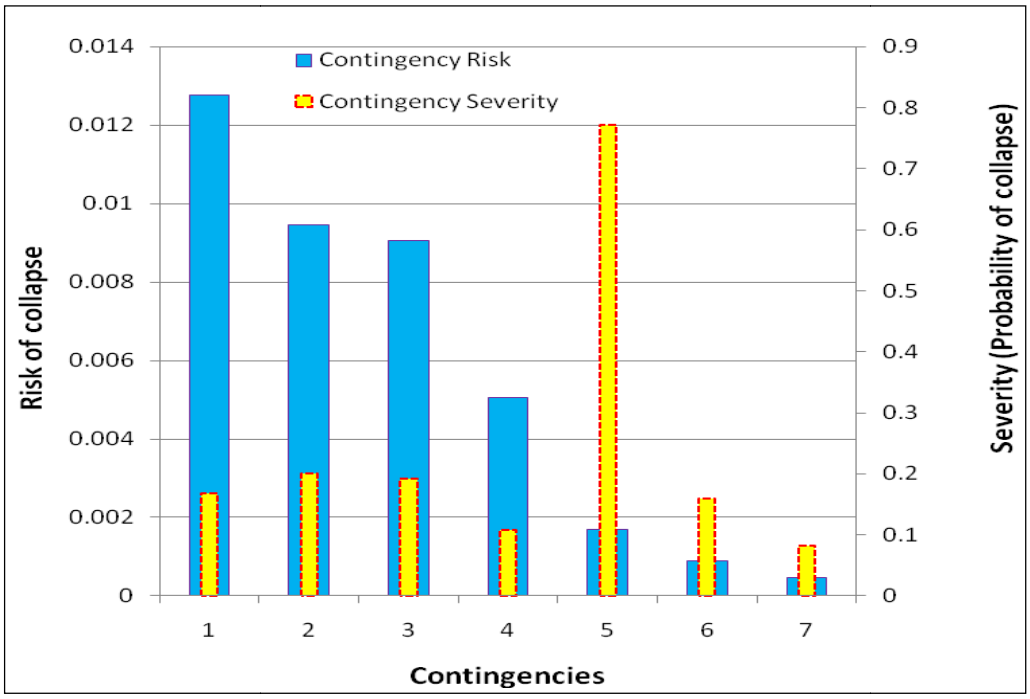


Fig. 4.9 Contingency severity and risk

We see that the proposed M/C based contingency risk estimation and ranking method works well in identifying the top contingencies. The top contingencies can be screened using a specific cut off value of risk (say, average risk), and decision rules can be derived for the screened critical contingencies posing significant risk over probable operating conditions.

4.4.1.5 Computational Benefit

The proposed risk based contingency ranking method saves a huge amount of computational cost since it uses linear sensitivities computed along multiple stress directions and utilizes machine learning method to estimate severity of every scenario. For instance, the

risk estimation of 7 contingencies, for a sampled 15 stress directions in the study, required 15*7 = 105 CPF simulations and linear sensitivity computations to estimate severity of every contingency over 975 different loading conditions, as shown in Table 4.5. If not for the linear sensitivities, the conventional method would require a huge computation of about 975*7=6825 CPF computations to compute the margin stability or 6825 dynamic simulations to compute dynamic performance.

Table 4.5 Computational benefit of proposed CRE

Case	Contingencies	Operating Conditions	Total simulations
Uncertainty: Loads			
Conventional	7	975	6825
Proposed CRE *(k=15)*	7	975	105
Uncertainty: Loads and SVCs			
Conventional ***(estimation)***	7	3900	27300
Proposed CRE *(k=15)* ***(estimation)***	7	3900	105

So the computational cost of proposed CRE doesn't even depend on the number of operating conditions sampled, but only on number of homothetic stress directions sampled. If a very few homothetic stress directions has the ability to effectively characterize the load state space, then the computational cost to estimate contingency severity is highly reduced, as shown in Table 4.5.

The proposed CRE method's ability to reduce computational cost drastically for contingency ranking is bolstered when we consider some discrete parameter uncertainties also, such as SVC unavailability or generator group unavailability etc, in the stage of Monte Carlo sampling of basecases. Table 4.5 shows the estimated computational requirements for conventional and proposed CRE method of contingency risk estimation for operational state space comprised of both loading conditions and 2 SVC unavailabilities. There could be 4

combinations of 2 SVC states, i.e., both unavailable (00), one of them unavailable (01 and 10) and both available (11). So systematically combining these 4 states with the sampled 975 loading conditions, we obtain 3900 basecases or operating conditions. So the conventional contingency severity estimation method will have to perform 3900*7 = 27300 simulations for 7 contingencies. But the computational requirements of the proposed CRE method based on linear sensitivities and machine learning still proportional only to the number of stress directions characterizing the load state space. The influence of discrete parameter, i.e., SVC unavailability states can be accounted using the linear sensitivities, i.e., the sensitivity of stability margin with respect to reactive power injection at the SVC buses, as was used successfully in chapter 3 to find the boundary region.

It should be noted that the proposed CRE I and II have both almost similar computational requirements, as shown in Table 4.6. So the proposed contingency risk estimation method enables tremendous computational cost reduction for the purpose of risk based contingency ranking of multiple contingencies over several operating conditions sampled. The number of stress directions sampled could be increased further for increased accuracy, and still the computational requirement would be very less compared to full-fledged conventional contingency simulations.

Table 4.6 Computational requirements of proposed CRE I and CRE II

Contingency = 1 and k = 15	CRE I	CRE II
CPF computations	15	15
Linear Sensitivity computations	15	15
Stress directions probability estimation using IBk	Yes	No
Stress directions mapping to operating conditions using IB*k*	No	Yes

4.4.2 Multiple Contingencies Security Assessment

4.4.2.1 Contingency Grouping

This section presents the results for the proposed contingency grouping concept. The following 5 contingencies have been considered: Cordemais bus bar fault, Flamanville unit-2 outage, Chinon unit outage, Launay bus bar fault, and Domloup bus bar fault.

Figure 4.10 shows the progressive entropy curves for all the above mentioned contingencies in the total Brittany load state space. The possible contingency group recommendations are as shown in Fig. 4.11. This is because of their closeness and their nature of progression along the operating conditions through various ranges of loads. So the proposed grouping promises reduction in the number of operational rules from five to two.

The training databases required for validating the group recommendations are four as shown in Fig. 4.12. Therefore the contingency grouping also promises computational cost benefit by reducing the number of training databases required from five to four. The two best common decision rules for all the five contingencies are finally selected by rule validation process using an independent test data.

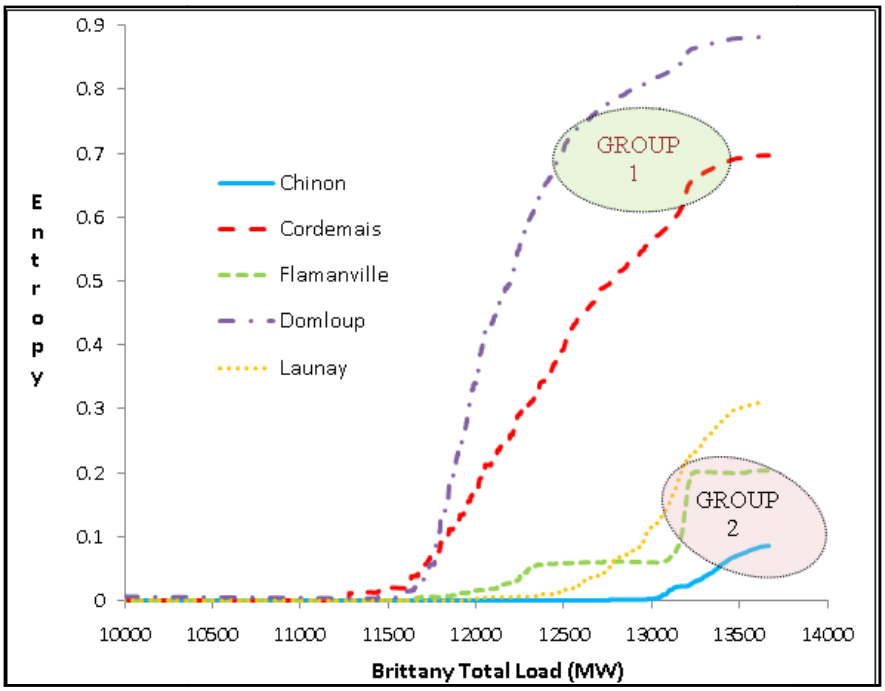


Fig. 4.10 Progressive entropy based contingency grouping

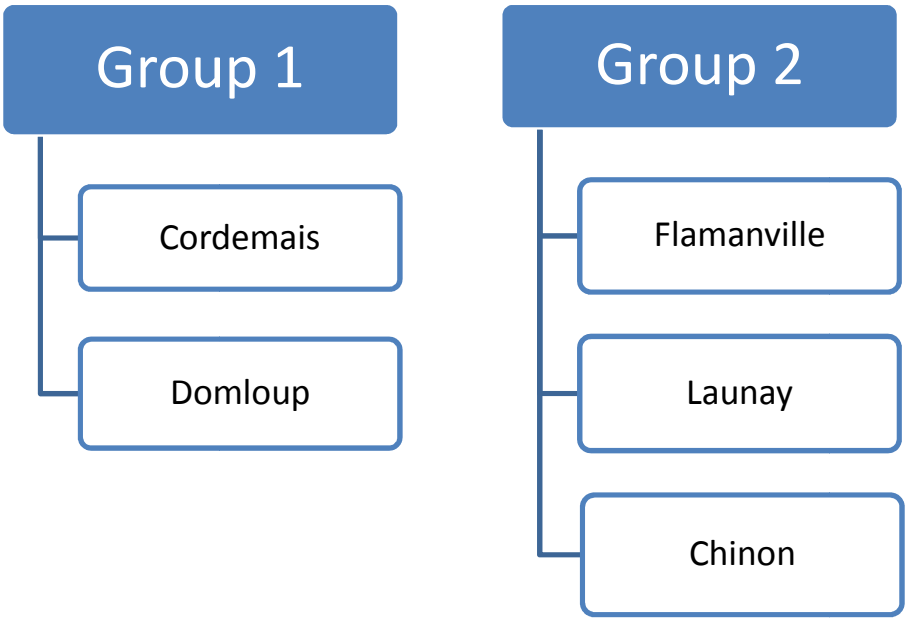


Fig. 4.11 Contingency Group Recommendations

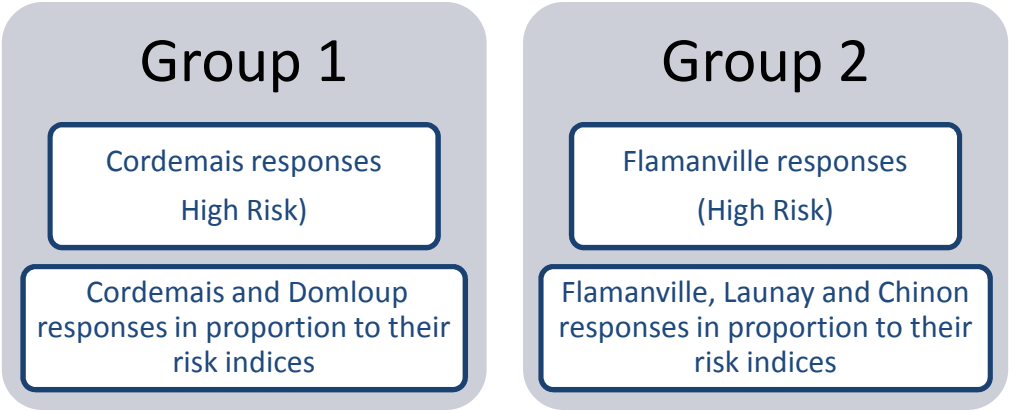


Fig. 4.12 Training Databases required to be generated

Figures 4.13, 4.14 and 4.15 show the progressive entropy curves of various contingencies on other variables, namely Cordemais bus voltage, total SEO region reactive reserve, and Chinon generator group reactive reserve.

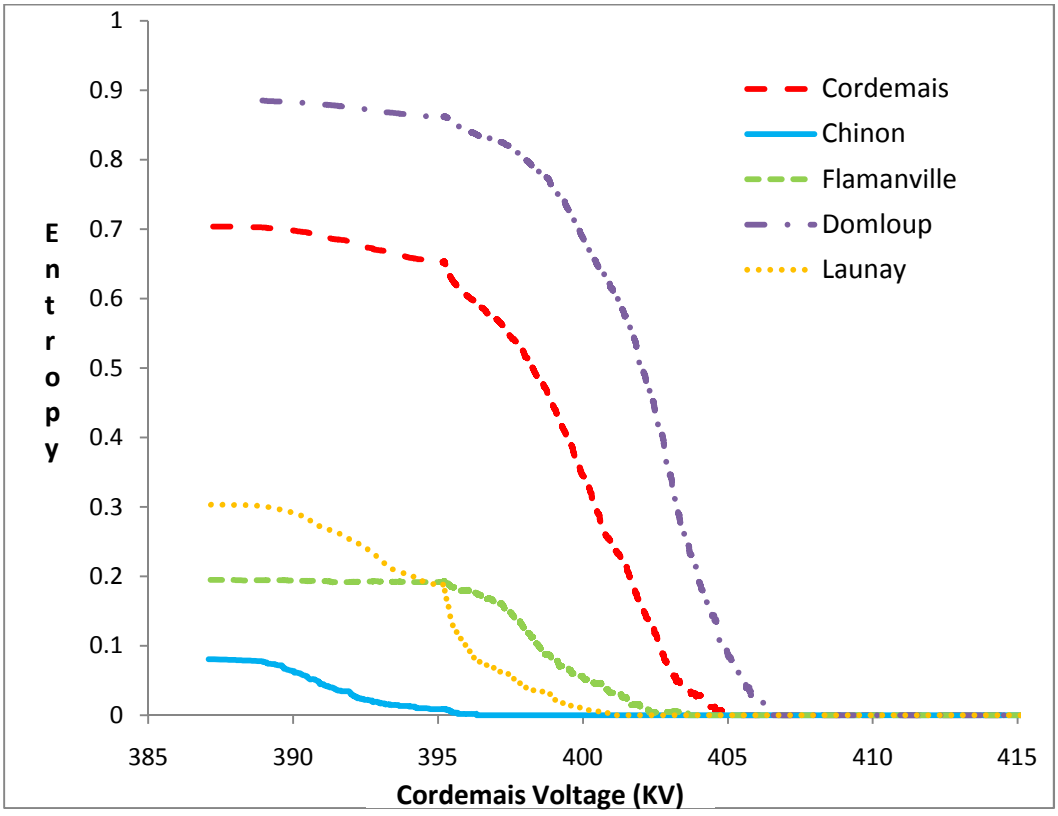


Fig. 4.13 Progressive entropy curves on Cordemais Voltage

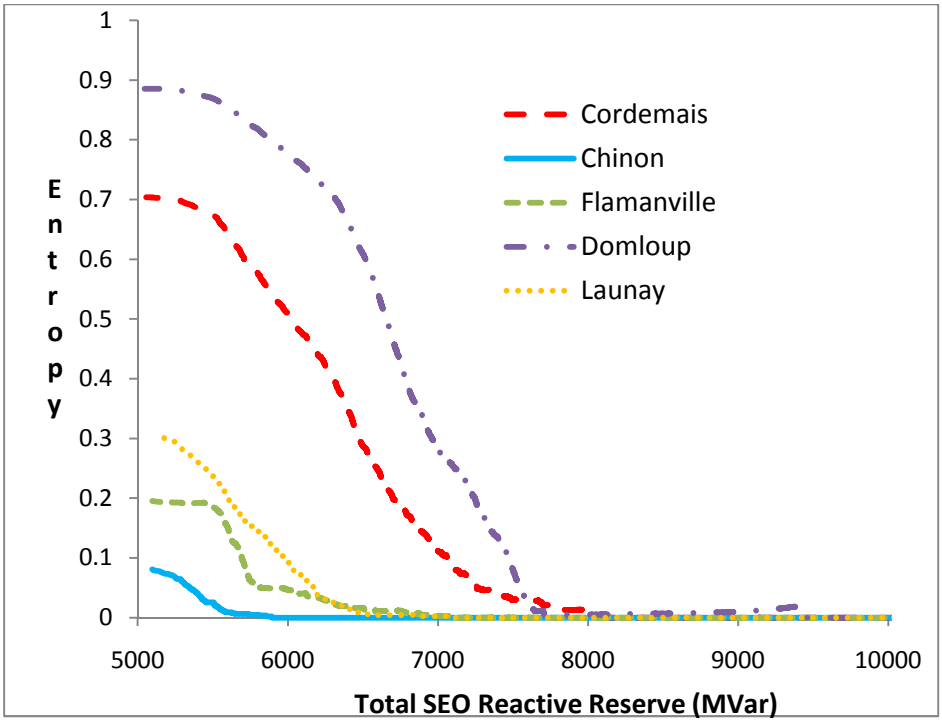


Fig. 4.14 Progressive entropy curves on total SEO reactive reserve

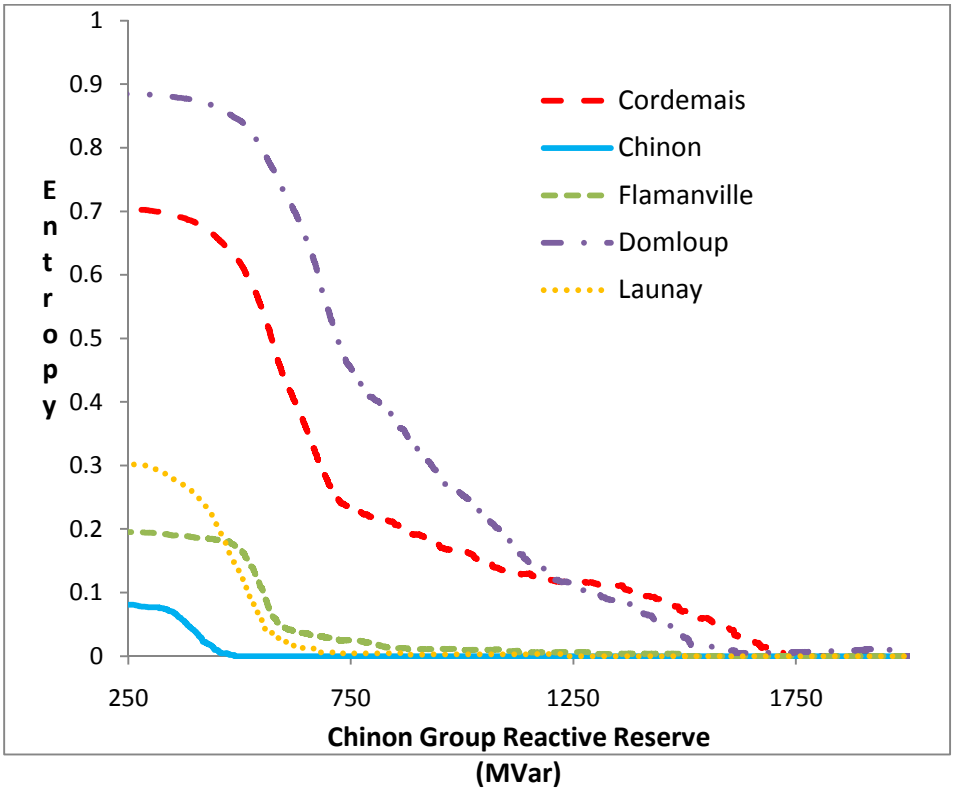


Fig. 4.15 Progressive entropy curves on Chinon group reactive reserve

The plots based on the above power system variables too produce similar contingency grouping recommendations, corroborating the recommendation based on the load variable. But the advantage of plotting the progressive entropy for variations in load parameter is that it is the sampling parameter, and using linear sensitivities the performance measures are computed without full-fledged simulation. So this saves a lot of computation, and promises further computational requirements savings at the stage of training database generation.

Figure 4.16 shows both the estimation and simulation output of progressive entropy curves for Cordemais contingency along the load variable. It was done for a sample of 975 loading conditions randomly selected form the multivariate loading distribution.

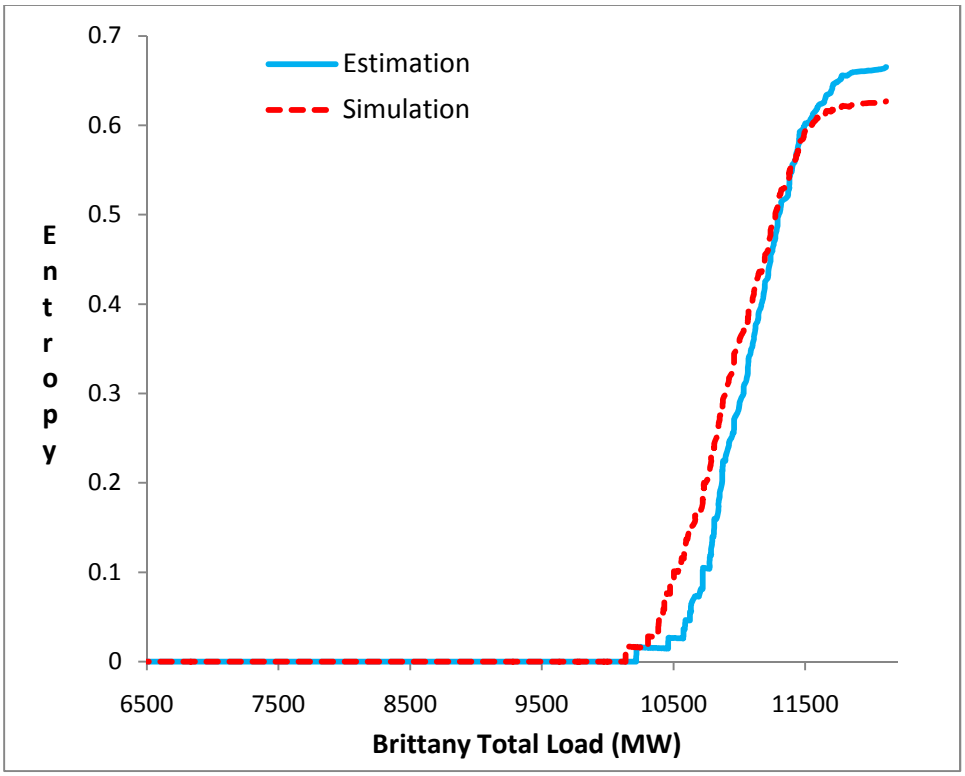


Fig. 4.16 Progressive entropy estimation vs. simulation

4.4.2.2 Operating Rules Validation

The study specifications for sampling the operating parameters, i.e., loading conditions, generators group unavailability, and SVCs unavailability in Brittany area, are similar to the study described in chapter 2, with a minor change regarding generator groups considered. In this study the main production units considered are nuclear groups in Civaux, Blayais, and St-Laurent. The units at Flamanville and Chinon are considered as part of the contingency, and so are not included in the sampling strategy. So the three units are sampled such that each of these three unavailabilities are represented in $1/4^{th}$ of the total basecases. The contingencies are applied at 900s and the ASTRE dynamic simulation is terminated at 1500s. The criteria used for labeling scenarios based on post-contingency responses are based on EHV bus voltages and simulation status at 1500s, same as chapter 2 specifications. Finally the training databases are formed, which contains 400KV voltages, SVC outputs and generator group reactive reserves sampled at 890s of simulation as the attributes and scenario labels as the class attribute.

The following results present the performances of various operating rules derived from a variety of training databases, including the databases recommended in the section 4.4.2.1 by the progressive entropy based contingency grouping method. Every training database is around the same size containing about 8000 operating conditions. Independent test databases are formed for every contingency separately by exactly following the same sampling and simulation specifications as mentioned above. All the independent test sets contains about 4000 instances.

Table 4.7 presents the performance results of rules for each contingency derived from separate a decision tree based on training database containing its respective post-contingency

responses. Rule for each contingency is tested against its respective test set. It can be seen that the classification accuracies for every contingency from separate decision trees are very high. But in this case, we end up with five separate rules for five contingencies.

Table 4.7 Separate operating rule for every contingency

S No	Contingency	Accuracy	FA	Risk
1	Cordemais	94.9783	0.034	0.122
2	Domloup	95.2081	0.039	0.068
3	Flamanville	99.3467	0.001	0.203
4	Chinon	99.3723	0.002	0.308
5	Launay	98.1378	0.008	0.19

Table 4.8 shows the result of rule performance when a common rule is derived from the training database containing only the contingency responses of Cordemais bus bar fault, the contingency with highest risk. The common rule is tested against the specific contingencies test data, and it is seen that the common rule based on Cordemais contingency doesn't perform well for all the other contingencies. For all the contingencies with lower severity than Cordemais, i.e., the contingencies at Flamanville, Launay and Chinon, the false alarms have increased tremendously. So a common rule based on worst case contingency alone will not be suitable for all the other contingencies, including Domloup which is grouped together with Cordemais for its similar severity levels at various load ranges as shown by progressive entropy curves in Fig. 4.10.

Table 4.8 One common rule based on Cordemais contingency responses

S No	Contingency	Accuracy	FA	Risk
1	Cordemais	94.9783	0.034	0.122
2	Flamanville	82.5067	0.174	0.203
3	Chinon	82.2067	0.18	0
4	Domloup	87.7057	0.011	0.388
5	Launay	87.2793	0.135	0

Table 4.9 shows a common operating rule formed by generating a training database with operating conditions containing post-contingency responses of every contingency proportional to its risk index, as shown in Table 4.4. We can see that the rule doesn't perform well for the most constraining contingency at Cordemais, apart from its poor performance for other contingencies too. So such a common decision tree requires meta-learning techniques to improve its accuracy further, at the cost of overfitting the tree and complicating the operating rule.

Table 4.9 One common rule based on all the contingency responses

S No	Contingency	Accuracy	FA	Risk
1	Cordemais	90.6	0.003	0.5
2	Flamanville	91.331	0.078	0.375
3	Chinon	90.82	0.09	0.273
4	Domloup	84.85	0	0.532
5	Launay	96.36	0.026	0.2

Table 4.10 shows the results for operating rule performance when Cordemais is grouped with other contingencies. Common operating rule is derived for each group based on training database containing contingency responses proportion to risk indices of contingencies in that respective group. It is seen that the recommended grouping of Cordemais contingency with Domloup contingency has the best performance, where the rule's performance for Cordemais in on par with the highest performance obtained in Table 4.7 and the rule's performance for Domloup betters the performance in Tables 4.8 and 4.9. The reduction in common rule's performance for Domloup contingency compared to Table 4.7 performance can be traded off against the fact that Domloup contingency has the least

risk index with very less probability and the prospect of reducing the number of operating rules atleast by one for operator's convenience. The rule could be further improved by increasing the representation of post-contingency responses of Domloup contingency more in the training database.

Table 4.10 Cordemais contingency grouped with other contingencies

Contingency	*Accuracy*		
Cordemais	94.8576	92.66	92.63
Domloup | Flamanville | Chinon	89.09	87.07	86.09

Fig. 4.17 shows the top five rule attributes for Group-1 contingencies, with stars placed at Cordemais and Domloup contingency locations.

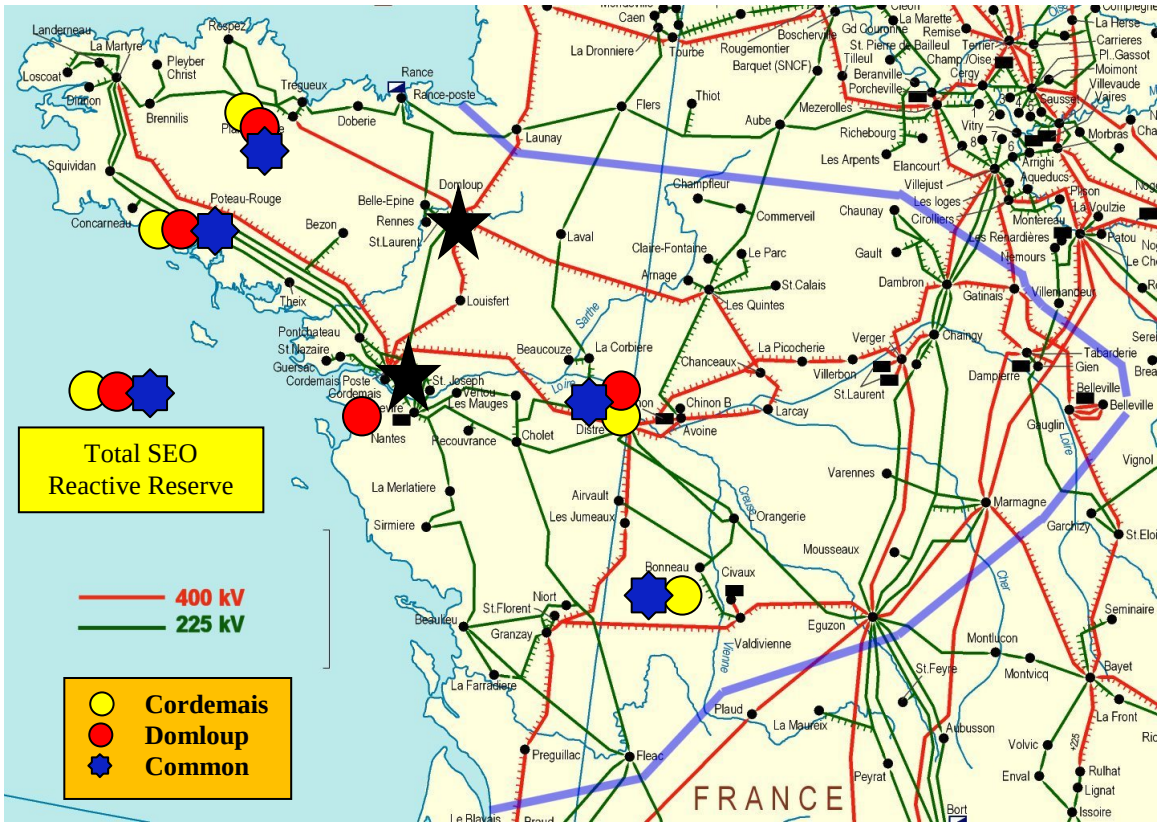


Fig. 4.17 Top five operating rule attributes for Group-1 contingencies

It shows rule attributes derived from Cordemais post-contingency response database, Domloup post-contingency response database and also the common training database produced based on proportional representation of both the contingency responses according to their risk indices. The commonality of the rule attributes for each case in the French Grid justifies the grouping of these two contingencies together for security assessment. Nevertheless, it should be noted that though the rule attributes are similar, the order they appear in the tree and their respective thresholds are different due to the differences in the training databases for each case.

Table 4.11 shows the results justifying the Group-2 recommendation made in section 4.4.2.1, and also aids in finalizing the common operating rule for the Group-2 contingencies. Along the columns is different training databases generated starting from a database made of Flamanville contingency responses only, then Launay contingency responses only, then Flamanville and Launay responses together according to the proportion of their their risk indices, and finally Flamanville, Launay and Chinon responses together according to the proportion of their their risk indices. The first **Flamanville** and fourth **Flamanville & Launay & Chinon** are the recommended training databases as per Group recommendation. So it can be observed that both the recommended training databases are producing operating rules that perform well. The rule from **Flamanville & Launay & Chinon** training database gives the best performance for all the contingencies, and the rule from **Flamanville** performs well in proportion to the contingency's risk index, i.e., for Flamanville with high risk the performance is the best and for Launay with the lowest risk the performance is least but still high enough.

Table 4.11 Group-2 contingencies rule performances from various training databases

Training Database / Contingency	*Flamanville*	*Launay*	*Flamanville & Launay*	*Flamanville & Launay & Chinon*
Flamanville	99.3467	92.86	97.67	97.1691
Launay	93.93	98.1378	94.6	95.1515
Chinon	96.517	95.437	97.465	98.1169

The conclusion is that:

1. The contingency grouping recommendation based on progressive entropy doesn't give importance to the proximity of contingencies on the French Grid, but is based on the similarity of contingency effects on the operating conditions along all the load ranges. The final grouping of contingencies is shown in Fig. 4.18.
2. The group recommendations guide in reducing the number of operating rules for operator's convenience and also in generating set of common rules with good performance for multiple contingencies. This is better than having a common rule for all the contingencies performance wise, and having separate operating rules for every contingency convenience wise.
3. The decision on best operating rule is taken based on the rule's performance on various contingencies within the group, weighed according to the risk levels of each contingency.
4. Even if the rules are to be improved by some feedback or meta-learning techniques, this is a better starting point as the degree of complexity of the final rule will be reduced.
5. By using linear sensitivities progressive entropy curves for all the contingencies

along the load variable are computed at much reduced computation, which further helps in reducing the computational requirements for generating training databases. This is achieved using the guidance obtained from the contingency grouping stage.

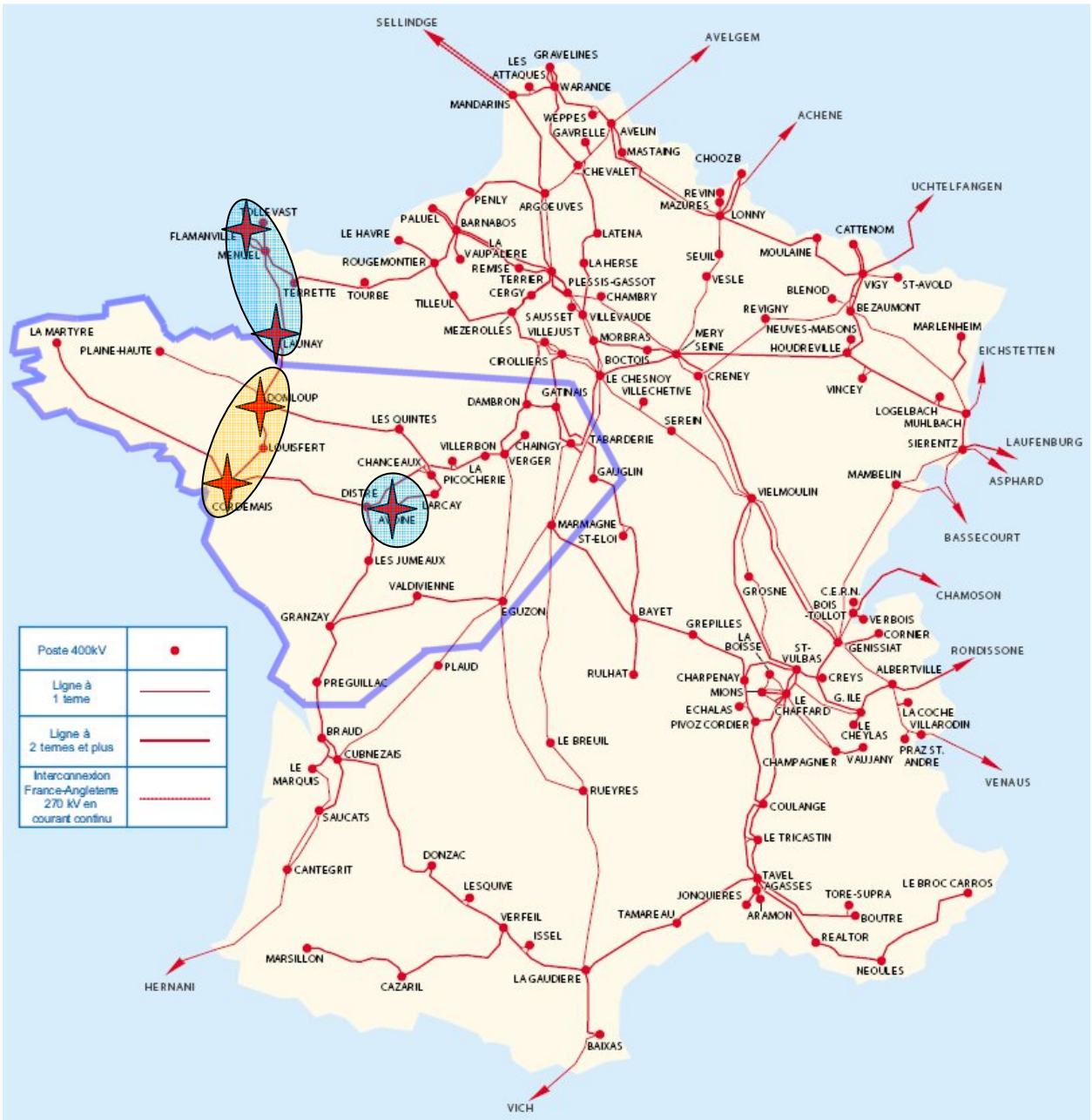


Fig. 4.18 French EHV network – contingency grouping recommendations

6. The proposed criteria of grouping contingency is mainly visual right now, but it can be advanced to include quantitative index by using machine learning techniques to find the closeness in multivariate regions.
7. The proposed contingency grouping based on overlap of boundary regions can also be used to group contingencies for other applications, such as reactive power planning problems, special protection schemes design for a group of contingencies, investigating interactions among various defense schemes etc.

4.5 CONCLUSIONS

This chapter proposed a comprehensive decision tree based power system operational planning for multiple contingencies. The foundation for the chapter was laid by earlier chapters, where the process of efficient training database generation is proposed and illustrated. In this chapter the main contribution was the proposal of risk based contingency ranking method and the progressive entropy based contingency grouping method. The developed concepts were demonstrated on the French network for five critical contingency locations. The contingency risk estimation method based on linear sensitivities and machine learning techniques for non-parametric operating conditions distribution proved to produce realistic results at a much reduced computational cost. The contingency grouping method guided in obtaining lesser number of operating rules that performs well for all the contingencies in the respective groups, thereby providing system operators the benefit of dealing with lesser number of rules.

CHAPTER 5 CONCLUSIONS

5.1 CONCLUSIONS

Our primary focus in this dissertation has been on power system operational planning using decision trees against voltage instability issues. The primary motivation of this work is from the fact that the performance of the operating rules derived from such machine learning algorithms in real time depends heavily on the quality of database used for training. Most of the work in decision tree based security assessment in power system has focused in improving the decision tree algorithm to obtain better classification performance from rules. While some works have made the crucial observation about the requirement of good training database, there has not been any work that has developed a systematic procedure to generate a training dataset that has the ability to capture the most important and realistic operating conditions having significant influence on the decision making. Also, the issues of generating operating rules for many contingencies, regarding the classification performance and system operators' convenience, have not been given enough attention.

So, in this dissertation we have developed efficient methods to process the system scenarios for generating high information contained database for training the decision trees. The method is constructed based on Monte Carlo Variance reduction techniques and has been systematically illustrated on a large scale realistic power network of French EHV grid with 5331 buses, with explicit focus on the West France, i.e., Brittany region that is prone to voltage collapse situations during winter periods due to heavy loading. The results showed significant improvement in the classification performances of the decision trees offering tremendous economic benefits, all at greatly reduced computational requirements inspite of

considering non-parametric multivariate distributions of operating parameters for sampling operating conditions. The results were analyzed in detail and the importance of generating such intelligent databases for training has been established.

The latter part of the dissertation developed a systematic approach to perform decision tree based security assessment of multiple contingencies. A risk based contingency ranking method based on instance based learning algorithm was developed, taking into consideration the non-parametric nature of operating conditions probability distribution. Also a contingency grouping method was proposed that enabled generating minimum number of well performing operating rules for many contingencies, with an idea to alleviate the burden for operators in making decisions.

All the reduction in computational requirements, i.e., in generating high information content training database, estimating risk indices for multiple contingencies, and also for generating operating rules for many contingencies, was achieved by the proposed Latin Hypercube Sampling of stress directions in multivariate state space, and also by the use of linear sensitivities of performance measures.

The specific contributions of the work in this dissertation are:

1. **Efficient processing of system scenarios:** An approach to efficiently sample system scenarios in machine learning studies for power system security assessment that increases classification accuracy while reducing computing requirements.
    a. *Sampling from correlated non-parametric multivariate distribution:* The non-parametric and dependence structure of expected loading scenarios, according to historical observations, were taken into account. This result in generating operating rules providing higher classification accuracy, more economic rules with

interesting monitoring locations that are closer to the contingency event.

b. *Fast state space characterization:* A Latin hypercube sampling of stress directions and linear sensitivities based method was developed for very fast identification of high information content region (boundary region) in the multivariate operating parameter state space. Since it is based on Monte Carlo simulation, it doesn't face any computationally intractable situations as some analytical methods may face in finding the closest boundary limits for large scale systems.

2. **Operational security rules of multiple contingencies:** A comprehensive methodology to perform decision tree based security assessment for multiple contingencies.

a. *Risk-based contingency ranking:* A risk-based contingency ranking method has been developed that helps in screening most critical contingencies for planning under a wide range of scenarios. The method gives accurate risk indices since it considers the realistic possibility of loading conditions following any likely stress directions from the non-parametric historical distribution. The computational cost involved in ranking many contingencies is greatly reduced by using linear sensitivities.

b. *Contingency Grouping:* A contingency grouping method based on newly devised metric called *progressive entropy* is developed that guides in generating the minimum number of well performing operating rules for all the contingencies, thereby benefiting system operators.

3. **Real-time application:** The developed methods are systematically implemented in French power network, focusing on the west France, Brittany region. The dissertation

provided solutions for a realistic voltage stability related operational planning problem that SEO region of French network faces every winter. The RTE-France company is on its way to apply the developed efficient processing methodology also for an investment planning problem this summer.

5.2 Future Work

Special Protection System (SPS) reliability assessment: The main difference between deriving operating rules and SPS logic are:

a. The SPS logic is automated.

b. The SPS logic is not only limited to critical operating condition detection with respect to some stability criteria, but also involves automatic corrective action to safeguard the system against impending instability.

Also, there are important questions to be answered regarding SPS's reliable operation from a 'system level view', such as:

(i) Are there system operating conditions (topology, loading, flows, dispatch, and voltage levels) that may generate a failure mode for the SPS?

(ii) Are there two or more SPS that may interact to produce a failure mode?

So the objective is to develop a decision support tool to perform SPS failure mode identification, logic re-design and risk assessment from a 'systems view'. The contingency grouping concept will be used to reduce the problem dimension in identifying the possible failure modes due to SPS interactions, thereby reducing the computational burden and analysis complexity. The efficient scenario processing method developed in the dissertation will be used to identify failure modes, estimate risk indices and re-design SPS logic.

REFERENCES

[1] Western Electricity Coordinating Council, (2003, Apr.), NERC/WECC Planning Strandards, [Online] Available: http://www.wecc.biz/documents/library/procedures/planning/WECC-NERC_Planning%20Standards_4-10-03.pdf

[2] A.M Abed, “WSCC voltage stability criteria, under voltage load shedding strategy, and reactive power reserve monitoring methodology,” *IEEE Power Engineering Society Summer Meeting*, vol. 1, pp. 191-197, 18-22 July 1999

[3] R. Billinton, L. Salvaderi, J.D. McCalley, H. Chao, Th. Seitz, R.N. Allan, J.Odom, and C. Fallon, ” Reliability issues in today's electric power utility environment,” *IEEE Trans. Power Systems*, Vol. 12, issue 4, pp. 1708-1714, Nov 1997

[4] M. J. Beshir, “Probabilistic based transmission planning and operation criteria development for the Western Systems Coordinating Council,” IEEE Power Engineering Society Summer Meeting, Vol.1, pp. 134-139, 18-22 Jul 1999

[5] A. A. Chowdhury, and D. O. Koval, “Probabilistic assessment of transmission system reliability performance,” IEEE Power Engineering Society General Meeting, 7 pp, 2006

[6] W. Li, and P. Choudhury, “Probabilistic Transmission Planning,” Power and Energy Magazine, IEEE, Volume: 5, Issue: 5, Sept.-Oct. 2007

[7] H. Wan, J. D. McCalley, and V. Vittal, “Risk based voltage security assessment,” IEEE Trans. Power Systems , Volume 15, Issue 4, Nov. 2000 pp.1247 - 1254

[8] F. Xiao, and J. D. McCalley, “Risk-Based Security and Economy Tradeoff Analysis for Real-Time Operation,” IEEE Trans. Power Systems , Volume 22, Issue 4, Nov. 2007

Page(s):2287 - 2288

[9] R. Billinton and W. Li, *Reliability Assessment of ElectricPower Systems Using Monte Carlo Methods*, New York: Plenum Press, 1994

[10] S. Henry, J. Pompee, L. Devatine, M. Bulot, and K. Bell, "New trends for the assessment of power system security under uncertainty," *IEEE PES Power Systems Conference and Exposition*, vol.3, pp. 1380-1385, 10-13 Oct. 2004

[11] A. C. G. Meio, J. C. O. Mello, and S. Granville, "The effects of voltage collapse problems in the reliability evaluation of composite systems," IEEE Trans. Power Systems, Volume: 12, Issue: 1, pp. 480-488, Feb 1997

[12] A. A. Chowdhury, L. Bertling, B. P. Glover, and G. E. Haringa, "A Monte Carlo Simulation Model for Multi-Area Generation Reliability Evaluation," Probabilistic Methods Applied to Power Systems, pp. 1-10, 11-15 June 2006

[13] M. J. Rosero, and M. A. Rios, "Characterization of the Maximum Loadability in Power Systems Due to Contingencies in the Operative Planning Scenario," Power Tech, 2007 IEEE Lausanne, pp. 1272-1277, 1-5 July 2007.

[14] J. D. McCalley, V. Vittal, and N. Abi-Samra, "An overview of risk based security assessment," Power Engineering Society Summer Meeting, IEEE, volume: 1, pp. 173-178 vol.1, 18-22 Jul 1999.

[15] J. McCalley, S. Asgarpoor, L. Bertling, R. Billinion, H. Chao, J. Chen, J. Endrenyi, R. Fletcher, A. Ford, C. Grigg, G. Hamoud, D. Logan, A. P. Meliopoulos, M. Ni, N. Rau, L. Salvaderi, M. Schilling, Y. Schlumberger, A. Schneider, and C. Singh, "Probabilistic security assessment for power system operations", Power Engineering Society General Meeting.

IEEE, pp. 212 - 220 Vol.1 6-10 June 2004.

[16] X. Yu, and C. Singh, “Probabilistic Analysis of Loadability in Composite Power Systems Considering Security Constraints,” Power Systems Conference and Exposition, pp. 832-838, Oct. 29 2006-Nov. 1 2006

[17] L. Wehenkel, “Automatic Learning techniques in power systems,” Kluwer Academic Publishers, 1998

[18] Ian H. Witten and Eibe Frank, (2000), Data Mining: practical machine learning tools and techniques with Java implementations, Morgan Kaufmann Publishers, San Fransisco, CA

[19] Q. Zhou, J. Davidson, and A. A. Fouad, “Application of artificial neural networks in power system security and vulnerability assessment,” IEEE Trans. Power Systems, Volume 9, Issue 1, pp. 525 – 532, Feb. 1994

[20] L. Wehenkel, “Machine learning approaches to power-system security assessment,” IEEE Expert, IEEE Intelligent Systems and Their Applications, Volume 12, Issue 5, pp. 60 – 72, Sept.-Oct. 1997

[21] G. Zhou, and J. D. McCalley, “Composite security boundary visualization,” IEEE Trans. Power Systems, Volume 14, Issue 2, pp. 725 – 731, May 1999

[22] T. Niimura, H. S. Ko, H. Xu, A. Moshref, and K. Morison, “Machine learning approach to power system dynamic security analysis,” IEEE PES Power Systems Conference and Exposition, pp. 1084- 1088, 10-13 Oct. 2004

[23] L. Wehenkel, M. Glavic, P. Geurts, and D. Ernst, ” Automatic learning of sequential decision strategies for dynamic security assessment and control,” IEEE Power Engineering Society General Meeting, 2006., 6 pp.

[24] R. E. Abdel-Aal, “Short-term hourly load forecasting using abductive networks,” IEEE Trans. Power Systems, Volume 19, Issue 1, pp. 164 – 173, Feb. 2004

[25] L. Wehenkel, M. Pavella, E. Euxibie, E, and B. Heilbronn, “Decision tree based transient stability method a case study,” *IEEE Trans. Power Systems*, Volume 9, Issue 1, pp. 459-469, Feb. 1994

[26] N. D. Hatziargyriou, G. C. Contaxis, and N. C. Sideris, “A decision tree method for on-line steady state security assessment,” *IEEE Trans. Power Systems*, Volume 9, Issue 2, pp. 1052-1061, May 1994

[27] S. Rovnyak, S. Kretsinger, J. Thorp, and D. Brown, “Decision trees for real-time transient stability prediction,” *IEEE Trans. Power Systems*, Volume 9, Issue 3, pp. 1417-1426, Aug. 1994

[28] R. Diao, V. Vittal, K. Sun, S. Kolluri, S. Mandal, and F. Galvan, “Decision tree assisted controlled islanding for preventing cascading events,” *IEEE/PES Power Systems Conference and Exposition*, pp. 1-8, 15-18 March 2009

[29] R. Diao, K. Sun, V. Vittal, R.J. O'Keefe, M.R. Richardson, N. Bhatt, D. Stradford, and S.K. Sarawgi, “Decision Tree-Based Online Voltage Security Assessment Using PMU Measurements,” *IEEE Trans. Power Systems*, Vol. 24, Issue 2, pp. 832-839, May 2009

[30] CART website: http://salford-systems.com/cart.php

[31] ANSWER TREE website: http://www.spss.com/answertree/index.htm

[32] ORANGE website: http://www.ailab.si/orange/

[33] WEKA website: http://www.cs.waikato.ac.nz/ml/weka/

[34] C. Lebrevelec, P. Cholley, J.F. Quenet, and L. Wehenkel, “A statistical analysis of the impact on security of a protection scheme on the French power system,” *International Conference on Power System Technology, Proceedings, POWERCON '98,* Volume 2, pp. 1102 - 1106, 18-21 Aug. 1998

[35] Y. Schlumberger, C. Lebrevelec, and M. De Pasquale, “Power systems security analysis-new approaches used at EDF,” *IEEE Power Engineering Society Summer Meeting,* Volume 1, pp. 147 – 151, 18-22 July 1999

[36] J. Pierre, C. Lebrevelec, and L. Wehenkel, “Automatic learning methods applied to dynamic security assessment of power systems,” *International Conference on Electric Power Engineering. PowerTech Budapest,* pp.180, 29 Aug.-2 Sept. 1999

[37] H. Martigne, P. Cholley, D. King, and J. Christon, “Statistical method to determine operating rules in the event of generator dropout on EDF French Guyana Grid,” *IEEE Power Tech Proceedings, Porto,* Vol. 1, pp. 5, 10-13 Sept. 2001

[38] J. Paul and K. Bell, “A Flexible and Comprehensive Approach to the Assessment of Large-Scale Power System Security Under Uncertainty,” *Proc. of the 7th International Conference on Probabilistic Methods Applied to Power Systems,* Naples Italy, September 2002

[39] S. Henry, J. Pompee, M. Bulot, and K. Bell, “Applications of statistical assessment of power system security under uncertainty,” *International Conference on Probabilistic Methods Applied to Power Systems,* pp. 914-919, 12-16 Sept. 2004

[40] S. Henry, E. Bréda-Séyès, H. Lefebvre, V. Sermanson and M. Béna, “Probabilistic study of the collapse modes of an area of the French network,” *Proc. of the 9th International*

Conference on Probabilistic Methods Applied to Power Systems, Stockholm, Sweden, June 2006

[41] P. Cholley, C. Lebrevelec, S. Vitet, and M. de Pasquale, "Constructing operating rules to avoid voltage collapse: a statistical approach," *International Conference on Power System Technology, Proceedings, POWERCON '98*, Volume 2, pp. 1468-1472, 18-21 Aug. 1998

[42] C. Lebrevelc, Y. Schlumberger, and M. de Pasquale, "An application of a risk based methodology for defining security rules against voltage collapse," IEEE *Power Engineering Society Summer Meeting*, Volume 1, pp.185-190, 18-22 July 1999

[43] S. Henry, C. Lebrevelec, and Y. Schlumberger, "Defining operating rules against voltage collapse using a statistical approach: The EDF experience," *International Conference on Electric Power Engineering, PowerTech Budapest 99*. pp. 30, 29 Aug.-2 Sept. 1999

[44] Y. Schlumberger, J. Pompee, and M. De Pasquale, "Updating operating rules against voltage collapse using new probabilistic techniques," *IEEE/PES Transmission and Distribution Conference and Exhibition: Asia Pacific.*, Volume 2, pp. 1139-1144, 6-10 Oct. 2002

[45] G. C. Oliveira, M. V. F. Pereira, and S. H. F Cunha, "A technique for reducing computational effort in Monte-Carlo based composite reliability evaluation," IEEE Trans. Power Systems, Volume 4, Issue 4, pp. 1309 – 1315, Nov. 1989

[46] S. R. Huang, and S. L. Chen, "Evaluation and improvement of variance reduction in Monte Carlo production simulation," IEEE Trans. Energy Conversion, Volume 8, Issue 4, pp. 610 – 620, Dec. 1993.

[47] C. Singh, and J. Mitra, "Composite system reliability evaluation using state space

pruning," IEEE Trans. Power Systems, Volume 12, Issue 1, pp. 471 – 479, Feb. 1997.

[48] D. Lieber, A. Nemirovskii, and R. Y. Rubinstein, "A fast Monte Carlo method for evaluating reliability indexes," IEEE Trans. Reliability, Volume 48, Issue 3, pp. 256 – 261, Sept. 1999

[49] P. Jirutitijaroen, and C. Singh, "Comparison of Simulation Methods for Power System Reliability Indexes and Their Distributions," IEEE Trans. Power Systems, Volume 23, Issue 2, pp. 486 – 493, May 2008

[50] C. Marnay, and T. Strauss, "Effectiveness of antithetic sampling and stratified sampling in Monte Carlo chronological production cost modeling," IEEE Trans. Power Systems, Volume 6, Issue 2, pp. 669-675, May 1991

[51] T. V. Cutsem, L. Wehenkel, M. Pavella, B. Heilbronn, and M. Goubin, "Decision tree approaches to voltage security assessment," *IEE Proceedings on Generation, Transmission and Distribution*, Vol 140, Issue 3, pp. 189-198, May 1993

[52] Y. Jacquemart, L. Wehenkel, and P. Pruvot, "Practical contribution of a statistical methodology to voltage security criteria determination," *Proceedings of the 12th Power Systems Computation Conference*, pp. 903-910, 1996

[53] L. Wehenkel, C. Lebrevelec, M. Trotignon, and J. Batut, "A probabilistic approach to the design of power systems protection schemes against blackouts," *Proceedings of the IFAC Symposium on Control of Power Plants and Power Systems, CPSPP97*, pp. 506-511, 1997

[54] T. E. Dy-Liacco, "Enhancing power system security control," *Computer Applications in Power, IEEE*, Volume 10, Issue 3, pp. 38-41, July 1997

[55] X. Jiantao, H. Mingyi, W. Yuying and F. Yan, "A fast training algorithm for support

vector machine via boundary sample selection," *Proceedings of the International Conference on Neural Networks and Signal Processing*, Vol.1, pp.20- 22, 14-17 Dec. 2003

[56] G. M. Foody, "The significance of border training patterns in classification by a feedforward neural network using backpropagation learning," *Int. J. Remote Sens.*, vol. 20, pp. 3549-3562, 1999

[57] I. Genc, R. Diao, V. Vittal, S. Kolluri, and S. Mandal, "Decision Tree-Based Preventive and Corrective Control Applications for Dynamic Security Enhancement in Power Systems," *IEEE Trans. Power Systems*, Issue: 99, 2010

[58] S. Greene, I. Dobson, and F. L. Alvarado, "Contingency ranking for voltage collapse via sensitivities from a single nose curve." IEEE Trans. Power Syst.14: 232-240, 1996

[59] N. Amjady and M. Esmaili, "Application of a new sensitivity analysis framework for voltage contingency ranking," *IEEE Trans. Power Systems*, Vol. 20, Issue 2, pp. 973-983, 2005

[60] N. Senroy, G. T. Heydt and V. Vittal, "Decision Tree Assisted Controlled Islanding," *IEEE Trans. Power Systems*, Volume 21, Issue 4, pp. 1790-1797, Nov. 2006

[61] B. D. Ripley, *Stochastic Simulation*, Wiley: New York, 1987

[62] R. A. Thisted, *Elements of Statistical Computing*, Chapman and Hal ltd., 1988

[63] X. Yu and C. Singh, "Expected power loss calculation including protection failures using importance sampling and SOM," *IEEE Power Engineering Society General Meeting*, Vol.1, pp. 206- 211, 6-10 June 2004

[64] E. A. Unger, L. Harn and V. Kumar, "Entropy as a measure of database information," *Computer Security Applications Conference, 1990, Proceedings of the Sixth Annual*, pp. 80-

87, 3-7 Dec 1990

[65] R. Y. Rubinstein, *Simulation and the Monte Carlo Method*, Wiley: New York, 1981

[66] A. Rencher, *Methods of Multivariate Analysis*, Wiley: New York, 1995

[67] Luc Devroye, *Non-Uniform Random Variate Generation*, Springer, Newyork, 1986

[68] J. E. Gentle, *Random Number Generation and Monte Carlo Methods*, Springer, Newyork, 1998

[69] ASSESS, TROPIC, METRIX website: http://www.rte-france.com/htm/an/activites/assess.jsp

[70] I. Dobson, and L. Lu, "New methods for computing a closest saddle node bifurcation and worst case load power margin for voltage collapse," *IEEE Trans. Power Systems*, Vol. 8, Issue. 3, pp. 905-913, Aug 2002

[71] G. Papaefthymiou, and D. Kurowicka, "Using Copulas for Modeling Stochastic Dependence in Power System Uncertainty Analysis," *IEEE Trans. Power Systems* ,Vol. 24, Issue 1, pp. 40 – 49, Feb. 2009

[72] S. Greene, I. Dobson, and F. L. Alvarado, "Sensitivity of the loading margin to voltage collapse with respect to arbitrary parameters," *IEEE PES winter meeting* Baltimore, 1996

[73] B. Long, and V. Ajjarpau, "The sparse formulation of ISPS and its application to voltage stability margin sensitivity and estimation," *IEEE Trans. Power Syst.*, pp. 944-951, 1999

[74] V. Krishnan, H. Liu, and J. D. McCalley, " Coordinated reactive power planning against power system voltage instability," *IEEE/PES Power Systems Conference and Exposition*, pp. 1 – 8, 15-18 March 2009

[75] T. Van Cutsem and C. Vournas, Voltage Stability of Electric Power Systems. Boston:

Kluwer Academic Publishers, 1998.

[76] C. W. Taylor, Power System Voltage Stability. EPRI Power System Engineering Series. New York: McGraw Hill, 1994

[77] H. Liu, L. Jin, J. D. McCalley, R. Kumar, V. Ajjarapu, and N. Elia, “Planning Reconfigurable Reactive Control for Voltage Stability Limited Power Systems,” IEEE Trans. Power Systems, Vol. 24, Issue. 2, pp. 1029-1038, 2009

[78] Security Mapping and Reliability Index Evaluation: Security Margin Analysis Calculator, EPRI, Palo Alto, CA: 2000.

[79] V. Ajjarapu, “Computational Techniques for voltage stability assessment and control”, Springer; 1 edition (December 28, 2006)

[80] V. Ajjarapu, and C. Christy, “The continuation power flow: A tool for steady state voltage stability analysis.” IEEE Trans. Power Syst.7: 417-423, 1992

[81] W.G. Wyss, K.H. Jorgensen, “A User’s Guide to LHS: Sandia’s Latin Hypercube Sampling Software,” *Sandia National Laboratories Report SAND98-0210*, Albuquerque, NM, 1998.

[82] W. Hormann, J. Leydold, and G. Derflinger, *Automatic Non-uniform Random Variate Generation*, Springer, Newyork, 2004

[83] F. Capitanescu, and T. Van cutsem, “Unified sensitivity analysis of unstable or low voltages caused by load increases or contingencies,” *IEEE Trans. Power systems*, Vol. 20, pp. 321-329, Feb 2005

[84] V. Krishnan, “Coordinated Static and Dynamic Reactive Power Planning against Power System Voltage Stability Related Problems”, MS Dissertation, Iowa State University, Ames,

IA 2007

[85] M. Pandit, l. Srivastava, and j. Sharma, “Fast voltage contingency selection using fuzzy Parallel self-organizing hierarchical neural network,” *IEEE Trans. Power Systems*, Vol. 18, Issue. 2, May 2003

[86] V. Vittal, W. Kliemann, Y. X. Ni, D. G. Chapman, A. D. Silk, and D. J. Sobajic, Determination of generator groupings for an islanding scheme in the Manitoba Hydro system using the method of normal forms,” IEEE Trans. Power Systems, Vol. 13, Issue. 4, pp. 1345-1351, 1998

[87] B. Lee, S. Kwon, J. Lee, et al., “Fast contingency screening for online transient stability monitoring and assessment of the KEPCO system”, *IEE Proceedings Generation, Transmission and Distribution*, Vol. 150, No. 4, pp. 399-404, 2003.

[88] D. Fischer, B. Szabados, S. Poehlman, “Automatic contingency grouping using partial least squares and feed forward neural network technologies applied to the static security assessment problem,” *Large Engineering Systems Conference on Power Engineering*, pp.84-89, 2003

[89] L. Sigrist, I. Egido, E. F. Sánchez-Úbeda, and L. Rouco, “Representative Operating and Contingency Scenarios for the Design of UFLS Schemes,” *IEEE Trans. Power Systems*, Vol.25, Issue.2 , pp. 906-913, May 2010

[90] T.Y. Hsiao, C.A. Hsieh, and C. N. Lu; “A risk-based contingency selection method for SPS applications,” *IEEE Trans. Power Sys.*, Vol. 21, issue. 2, pp. 1009 – 1010, 2006

[91] C. P. Robert, and G. Casella, “Monte Carlo Statistical Methods,” 2004, Sringer, Newyork

[92] A. F. Atiya, “Estimating the posterior probabilities using the *k*-nearest neighbor rule,” *Neural Comput.*, vol. 17, no. 3, pp. 731–740, 2005

[93] L. Jiang, H. Zhang and Jiang Su, Learning *k*-Nearest Neighbor Naive Bayes for Ranking, Book chapter, Advanced Data Mining and Applications, ISBN: 978-3-540-27894-8, 2005